Birkbeck
UNIVERSITY OF LONDON

ORBIT - Online Repository of Birkbeck Institutional Theses

Enabling Open Access to Birkbeck's Research Degree output

The fully automated construction of metabolic pathways using text mining and knowledge-based constraints

https://eprints.bbk.ac.uk/id/eprint/40135/

Version: Full Version

Citation: Czarnecki, Jan Michael (2015) The fully automated construction of metabolic pathways using text mining and knowledge-based constraints. [Thesis] (Unpublished)

© 2020 The Author(s)

All material available through ORBIT is protected by intellectual property law, including copyright law.
Any use made of the contents should comply with the relevant law.

Deposit Guide
Contact: email

Thesis submitted for the degree of Doctor of Philosophy

The fully automated construction of metabolic pathways using text mining and knowledge-based constraints.

Jan Michael Czarnecki

Birkbeck, University of London

3rd May, 2015

I, Jan Michael Czarnecki, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I would like to thank my principal supervisor Dr. Adrian Shepherd for his unerring support and guidance over the past four years, and to my co-supervisors Dr. Irilenia Nobeli and Adrian Smith, at Unilever, for their invaluable insight into my project.

I would also like to thank my thesis committee chair Dr. Richard Hayward, and my other collaborators at Unilever, Dr. Gordon James and Dr. Diana Cox for their input from a priceless non-bioinformatics perspective.

Abstract

Understanding metabolic pathways is one of the most important fields in bioscience in the post-genomic era, but curating metabolic pathways requires considerable man-power. As such there is a lack of reliable experimentally verified metabolic pathways in databases and databases are forced to predict all but the most immediately useful pathways by inheriting annotations from other organisms where the pathway has been curated. Due to the lack of curated data there has been no large scale study to assess the accuracy of current methods for inheriting metabolic pathway annotations.

In this thesis I describe the development of the Literature Metabolic Pathway Extraction Tool (LiMPET), a text-mining tool designed for the automated extraction of metabolic pathways from article abstracts and full-text open-access articles. I propose the use of LiMPET by metabolic pathway curators to increase the rate of curation and by individual researchers interested in a particular pathway.

The mining of metabolic pathways from the literature has been largely neglected by the text-mining community. The work described in this thesis shows the tractability of the problem, however, and it is my hope that it attracts more research into the area.

Contents

List of Figures

List of Tables

Part I

Introduction

1 The Problem

PubMed currently contains over 24 million article records, and this number is increasing at a faster rate than ever with 2014 set to be the first year with over one million new article records [2]. In some fields, researchers are encouraged, or even required, to submit results to databases in a standard format. For instance, upon solving the structure for a protein, an X-ray crystallographer will submit the structure to the Protein Databank (PDB) as well as submitting the results in a paper for peer review [3]. This allows anybody with an Internet connection to find curated structures of proteins of interest quickly and efficiently. The data, being in a standard format, is also easily consumed by computer programs allowing large scale studies involving many structures to be carried out[1]. For instance, the CATH project has developed a semi-automated system for classifying protein domain structures through the comparison of structures in the PDB [4].

Unfortunately, this method of submitting results in a standard, computer-readable language is found in few fields in bioscience due to the breadth of fields and the pace with which new experiments can develop. We are currently in the situation where the vast majority of data in many fields is only available as unstructured text spread across many publishers' websites.

The study of metabolic pathways is one such field that suffers from a lack of manually curated data in databases. BRENDA is a large database of curated metabolic reactions, but individual reactions are not linked together to form pathways (meaning that there is little motivation to curate complete pathways from single organisms) [5]. KEGG [6] and BioCyc [7], are two databases that were developed to curate metabolic pathways. Ultimately, however, the databases are populated by human curators, which means it is practically impossible to keep up with all new articles being published. Training of a FlyBase Genetic Literature Curator, for instance, can take 6 months in addition to the time taken to actually curate an article [8].

[1] This may be a slightly optimistic view of PDB files, which are notorious for inconsistencies with the specification and requiring human intervention in parsing. While the format has been regularly updated to accommodate increasingly complex data, it was first conceived in the 1970s and many aspects, such as a fixed width of 80 characters (the width of a computer punch card), can cause difficulties in parsing files. Formats, such as PDBML/XML and mmCIF, have been developed to solve these problems.

In this thesis I will describe the development of LiMPET, the Literature Metabolic Pathway Extraction Tool. LiMPET is an attempt to speed up the discovery of metabolic pathway information in the literature and to aid curation through the use of text mining.

2 Metabolic Pathway Prediction

Because of the high cost (both economic and temporal) of manually curating documents, KEGG and BioCyc, two of the largest metabolic pathway databases, not only curate metabolic pathway data, but also predict metabolic pathways where there is no curated data. Both have different philosophies towards metabolism and how pathways are predicted.

2.1 BioCyc

BioCyc contains 3 database tiers. The tier 1 databases have received at least one year of manual literature-based curation, while databases in tiers 2 and 3 contain metabolic pathways predicted computationally from an organism's annotated genome sequence. Tier 2 databases have undergone a moderate amount of review, whereas tier 3 databases have had no manual review.

Tier 1 contains six organism databases — including EcoCyc, AraCyc and YeastCyc for the model organisms *Escherichia coli K-12 MG1655*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*, respectively. Tier 1 contains another database, MetaCyc, containing all metabolic reactions, from all organisms, for which experimental evidence in the literature has been curated. For instance, MetaCyc contains the pathway "vindoline and vinblastine biosynthesis" from the organism *Catharanthus roseus*. While *C. roseus* is a tier 3 organism, this pathway is of particular interest in the development of chemotherapy drugs and a large amount of research has been carried out on it. Therefore, the pathway is fully curated and belongs in tier 1. Many MetaCyc pathways even contain reactions from multiple organisms.

The method used to transfer annotations from tier 1 databases to lower tiers to form pathway predictions is well documented [9]. Computational annotations are obtained from the Comprehensive Microbial Resource and UniProt. MetaCyc pathways are used as reference pathways of small molecules connected by enzymes with specific E.C. numbers. Potentially expressed enzymes are identified from an organism's annotated genome and are assembled into pathways by matching their E.C. numbers with those in the MetaCyc reference pathways.

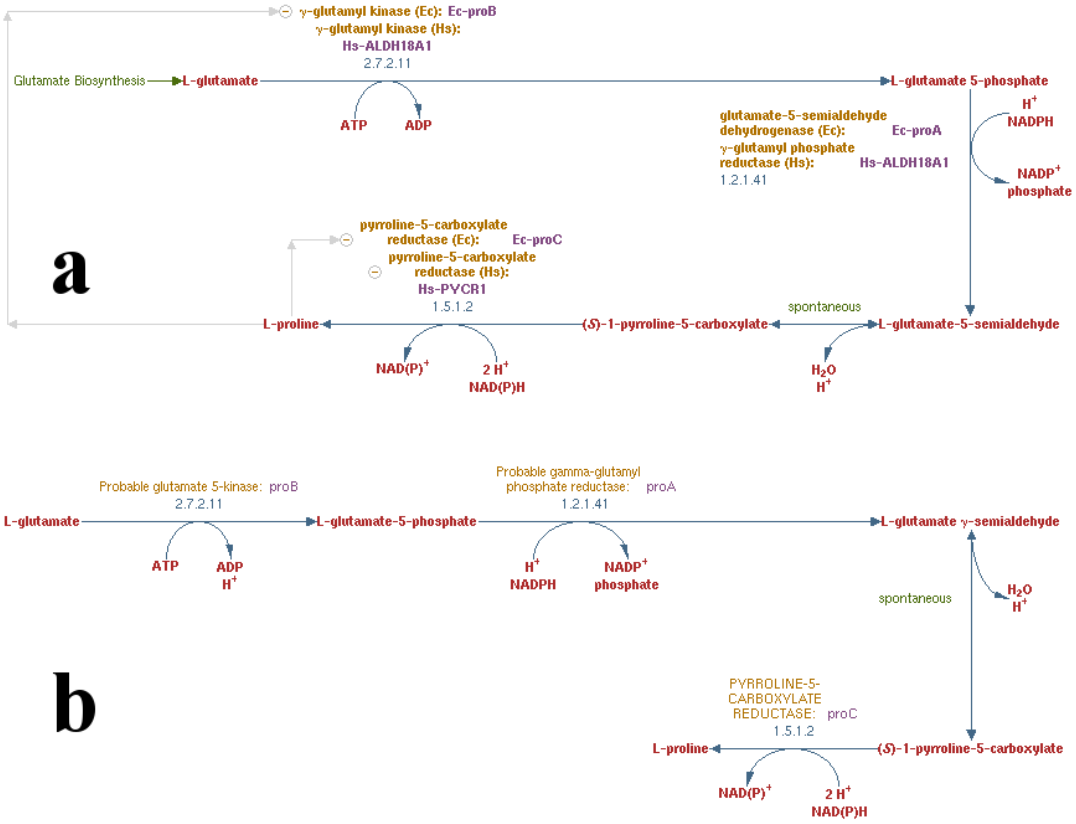

Figure 1: a) The "proline biosynthesis I" pathway from MetaCyc, built using curated metabolic reactions from *Escherichia coli* and *Homo sapiens*, b) The predicted "proline biosynthesis I" pathway for *Mycobacterium tuberculosis H37Rv*, built using enzyme predictions from the organism's genome sequence and using the MetaCyc pathway as a template.

Figure 1 shows an example of this annotation transfer.

BioCyc pathways are quite small and are generally used to show a specific process, such as the biosynthesis or degradation of a particular molecule. It is reasoned that the small pathways in BioCyc represent conserved biological processes and annotations can, therefore, be accurately transferred from one organism to another.

Paley and Karp carried out a small study into the accuracy of predictions made by the Pathway Tools suite when applied to *Helicobacter pylori* [10]. Of 98 predicted pathways, 40 showed evidence of their existence in the literature. It proved difficult to assess the accuracy of the tool, however, as the pathways which had no mention in the literature are not necessarily incorrect, but may simply be uncharacterised. It was deemed likely that 26 predicted pathways not found in the manual analysis of the literature, were in fact correct.

2.2 KEGG

KEGG has a different philosophy towards metabolism compared to BioCyc. Whereas BioCyc constructs small pathways, which have been determined to represent a single evolutionary block, KEGG constructs large networks which incorporate many BioCyc-sized pathways, showing how they interact with one another. Appendix I shows the KEGG pathway "Arginine and proline metabolism", which contains reactions corresponding to the MetaCyc "proline biosynthesis I" pathway shown in figure 1a as well as a host of other MetaCyc pathways.

The data model within KEGG is not as clear as that of BioCyc. Reference pathways are very general and all organisms and pathway variations are incorporated into a single network. There is little consistency in the organisms used to construct these reference pathways with lesser known organisms often being used. This evidence used to construct the reference pathways is kept apart from the separate organism databases. For instance, evidence from an *E. coli* experiment may have been used in the construction of a reference pathway. When viewing the individual reactions in this reference pathway the evidence will be displayed to the user (along with evidence from any other organisms). When looking at the pathway specifically in *E. coli*, however, no evidence will be visible.

2.3 Possible prediction inaccuracies

Unfortunately, it is currently impossible to properly assess the accuracy of the predicted pathways in either database. In order to carry out such an assessment we would require a large number of curated pathways from a range of organisms, which we, of course, do not have. This lack of curated pathways also means that there has been no large scale studies of metabolic evolution. Mano *et al.* developed a methodology for the alignment of metabolic pathways from different organisms, but found the scope of the study limited by the lack of curated pathways in multiple organisms within MetaCyc [11].

I carried out an initial study comparing the pathway predictions in a number of BioCyc tier 3 databases with their counterparts in KEGG. The E.C. numbers of enzymes predicted to be present in a number of BioCyc tier 3 pathways in multiple organisms were identified and compared to their presence in KEGG. Due to KEGG's licensing (described in Section 2.4), automating this comparison was not possible, but a manual comparison across a small selection of organisms identified inconsistencies between the two databases. Table 1 shows the predicted

Enzyme	*C. glutamicum*		*C. jeikeium*		*S. aureus*		*S. epidermidis*		*P. acnes*	
	BioCyc	KEGG	BioCyc	KEGG	BioCyc	KEGG	BioCyc	KEGG	BioCyc	KEGG
5.1.2.3										
6.2.1.3										
5.3.3.8										
1.3.99.3										
4.2.1.17										
1.1.1.35										
2.1.3.16										

Table 1: A table showing the enzymes of the MetaCyc pathway "fatty acid β-oxidation I" predicted to be present by BioCyc and KEGG in a number of organisms. A green cell specifies that a particular enzyme is predicted to be present, while a red cell specifies its absence.

presence of seven enzymes from the MetaCyc pathway "fatty acid β-oxidation I" in five organisms. While there are many differences between the predictions made by the two databases, it is impossible to determine which is correct in each case, due to the lack of evidence in the literature. Frustratingly, neither BioCyc nor KEGG show the steps made in making a specific prediction.

While still tremendously useful in helping researchers make hypotheses regarding unexplored pathways, the usefulness of these predicted metabolic pathways is somewhat diminished by the inability to determine how accurate they are. In Part V I show how LiMPET can be used to aid curation by using the tool to find evidence in the literature corroborating, or even contradicting, predicted pathways in BioCyc. While the curation of a single pathway would aid the research of anybody interested in said pathway, the curation of many pathways could lead to larger studies comparing pathways from many species which could feedback and improve prediction methods.

There are a number of different aspects of a given metabolic pathway that could be investigated by a larger study. Are there particular nodes that are more variable than others? How closely related must organisms be in order to reliably transfer annotations? Are some reactions more closely linked to the environment of the organism than others?

Linking pathway nodes to the host's environment has been investigated in non-metabolic protein interaction pathways. By comparing the known phylogenetic relationships and responses to certain stresses, Nikolaou *et al.* found that stress signalling pathways in fungi have evolved niche activity independent of phylogeny [12]. All organisms that were chosen for the study had fully sequenced and annotated genomes, allowing the evolutionary conservation of stress response genes to be studied. Three different stress response pathways (osmotic, oxidative and cell-wall stresses) were compared in 18 different fungal species which lived in a range of environments (human pathogens, plant pathogens and benign). For instance, *C. glabrata* (a human pathogen) was much more resistant to all three stresses than *S. cerevisiae* despite both being in the same family (*Saccharomycetaceae*).

Moreover, the study found that the core components of the stress response pathways (which are usually involved in other, non-stress response, pathways) were strongly conserved across the multiple species, while the upstream sensors and downstream transcriptional regulators showed a lower level of conservation. One might hypothesise that a similar pattern may occur in metabolic pathways, with core reactions being present in distantly related species and other reactions being more variable and evolving to fit a specific niche. One gap left by this study, however, is that the pathways were not experimentally verified. The study involved the comparison of genome sequences to determine pathway similarity. Just because a particular gene is present does not mean that a protein is expressed in a specific compartment at a specific point in the cell cycle, however. The pathway could, at least partially, take a different route around the protein.

Studies concerned with comparing metabolic pathways take the same approach. For instance, Huynen *et al.* compared the presence of TCA cycle genes in 19 sequenced genomes, finding distinct incomplete cycles that are linked to particular environments [13]. It was conceded, however, that that there was very little conservation in regulatory sequences and there was no way of predicting whether predicted genes were co-expressed.

Likewise, Gianoulis *et al.* attempted to quantify the environmental adaptation across a number to marine microbial organisms [14]. The group used data from the Global Ocean Survey ("a collection of quantitative environmental features and metagenomic sequences from more than 40 different aquatic sites") to map certain environmental features, such as temperature, salinity and depth, to metabolic features. Instead of mapping environmental features to

metabolic features on a one-to-one basis, they were linked with many-to-many relationships to create a "metabolic footprint". Some expected links where found. For instance, photosynthetic modules were found to be correlated with the environment, while the module for the ATP synthase complex, which is independent of the source of the energy, is abundant in all environments.

Despite the insight that can be gained from the comparisons of pathways predicted from genomic sequences, it is clear that experimentally verified data is needed to produce confident results. One of the principal aims at the outset of development of LiMPET was to create a tool that would aid in the discovery and curation of metabolic pathway data in the literature.

2.4 Funding of manual curation

In recent years the sustainability of bioinformatics databases has been questioned. Due to the worldwide recession, government funding for science has dropped and priorities in science funding have changed — leading to the termination, or commercialisation, of many databases. For instance, prior to 2013, The *Arabidopsis* Information Resource (TAIR) was the recognised source of curated sequence data from *Arabidopsis thaliana* for over a decade [15]. Despite widespread use and highly cited papers, in 2013 the US National Science Foundation declined to renew the project's funding and the database was forced into a subscription model [16].

Over the course of this project the effects of the high cost of manual curation has also become apparent in metabolic pathway research. While BRENDA has provided a commercial version of their database for some time, at the beginning of this project the whole database was downloadable in a flat file format [5]. While poorly maintained (the file format contained many inconsistencies that hindered parsing), I was able to undertake an initial analysis of the database. Unfortunately, free users of the database can no longer bulk download data or access the SOAP web services [17][2].

In 2011, due to significant cuts in government funding, KEGG introduced a subscription model for FTP access to their data [16]. While the data is still freely available through its website, large scale studies of many pathways are impossible without the ability to bulk download data. Then, on June 25th, 2014, it was announced (by Peter Karp on the BioCyc mailing list)

[2]The BRENDA maintainers have, however, released the BKM-React database [18] of cross-referenced BRENDA, KEGG and MetaCyc metabolic reactions. While continuing no species-specific BRENDA data, the database provided a collection of experimentally verified metabolic reactions that was used in this project to assess the correctness of extracted reactions.

that the funding for BioCyc had not been renewed. While the BioCyc team are reapplying for funding its situation is certainly precarious.

In the development of LiMPET I have investigated the semi-automated curation of metabolic pathways using text-mining, with the view of lowering the man-power and economic costs of curation and allowing individual researchers to find metabolic pathways in the literature themselves.

3 Text-mining

Computer processors are designed to follow very strict commands. This is reflected in the languages that are used to command computers, which, while incredibly varied, ultimately come down to providing a list of instructions for the various pieces of hardware within the computer. Therefore, the data which a computer program is designed to read and process must be stored in a strict format which the program can follow strict instructions to parse. Computer hardware, programs and data storage formats are all designed from the bottom up with this philosophy in mind. Natural language, however, isn't designed, but is constantly evolving and rules can often be hard to define. The English language is particularly notorious for having significant exceptions to the majority of spelling and grammatical rules.

Text-mining was first developed as a method in the field of business intelligence (BI) [19]. BI is concerned with the automated analysis of unstructured, often private, data available to a company to identify new business opportunities. As the World Wide Web grew over the 1990s, so too did the amount of freely available natural language text. Text-mining was soon of interest to any field were there was relevant text on the Internet. In science, it became standard to publish articles online, in addition to printed journals, and the potential of text-mining to aid manual curation in bioscience was recognised [20].

Text-mining development in bioscience initially focused on systems for named entity recognition (NER), the process of classifying elements in text into predefined categories, such as the names of proteins or small molecules[3]. Current state-of-the-art NER tools (focusing on entities such as proteins, small molecules, drugs and organisms) are able to achieve very high levels of accuracy — typically with F-scores (see Section 8 for an explanation of F-scores) greater than

[3] As I have made extensive use of publicly available NER tools in this project, I will describe their development in depth in Section 9.3.

90%. Focus has, therefore, shifted to interaction extraction, the process of determining the nature of relationships between different named entities. Interaction extraction can be used to determine abstract relationships, such as gene-disease relationships, or more direct, physical relationships, such as protein-protein interactions. As metabolic reactions fall into the latter category and the extraction of protein (and/or gene) interactions is the topic upon which most research has focused, a review of protein-protein interaction extraction methods provides a useful backdrop for the development of an extraction method for metabolic reactions.

3.1 Protein-protein interaction (PPI) extraction

There are a range of experimental methods that have been developed to characterise PPIs ranging from narrow focused methods such as X-ray crystallography, which offers the most convincing evidence that two proteins form a stable complex, to broad scoped methods such as yeast two-hybrid screens, which can find potential binding partners from a large pool of proteins. The IntAct database [21] (which contains curated PPIs from the 14 members of the IMEx Consortium [22]) contains interactions extracted from almost 13 000 publications (as of August 2014). While this is a monumental manual effort, it is still only a small fraction of the available material.

PPI extraction was the subject of one task at BioCreative II [23] in 2006, where teams were tasked with extracting PPIs from documents curated by IntAct and MINT (which were separate databases at the time before merging in the IMEx Consortium). Extracted interactions could then be compared to the gold standard, manually curated interactions. The best performing tool achieved an F-score of 29%, far lower than the high performance achieved by NER tools. Two general approaches to the problem were identified in the subsequent analysis of the submitted tools — which were termed as local association analysis and global association analysis [23]. Local association analysis identifies co-occurring proteins at either the sentence or passage level and may use other approaches such as interaction word lists and/or machine learning techniques to determine if an interaction between the co-occurring pair is described. Global association analysis focuses less on the characteristics of individual sentences, but rather looks at the co-occurrence of protein names multiple times in a document or over the whole collection. Global association analysis is more suitable for extracting well-known interactions that are described frequently in the literature, but only local association analysis is able to de-

termine novel interactions that have only been described once. The system developed during this project incorporates both local and global association analysis.

Kabiljo *et al.* [1] carried out a comparison on a range of PPI extraction tools including AkanePPI [24], OpenDMAP [25] and Whatizit [26]. AkanePPI is a state-of-the-art tool that utilises many natural language processing (NLP) methods. OpenDMAP is a general purpose information extraction platform which uses a heuristic approach. The patterns for PPI recognition were created manually to adapt the tool to the task. Whatizit is a suite of tools that can perform many bioscientific NLP tasks. The PPI extraction tool in Whatizit, Protein Corral, uses three methods which utilise co-occurrence and heuristic techniques.

A simple baseline method was also developed for the comparison. The method was co-occurrence based, looking for two protein or gene names within the same sentence as well as an "interaction" verb, such as *binds* or *phosphorylates* (a manually curated list of "interaction" verbs was used), in between the two entities — a similar methodology to the Co3 method of Protein Corral. The tools were evaluated on five gene-protein interaction corpora. While performance across the five corpora by each tool was variable, the simple baseline method showed an overall performance that was comparable to the more sophisticated methods, while being far simpler. I followed this simple methodology in the development of a metabolic reaction extraction method in Part III.

BioCreative III [27] proposed a slightly different PPI extraction task to that in BioCreative II. The task required the development of a tool capable of classifying and ranking abstracts according to their suitability for manual curation of PPIs in the full text. This behaviour is required by PPI databases, such as IntAct, to effectively manage their curator man-hours and to prevent the needless curation of irrelevant articles. Semi-automated selection of articles for manual inspection is common across the majority of biological annotation databases, but is typically carried out using simple PubMed searches. While effective at selecting articles relevant to a particular entity, this method is inadequate when dealing with complex events and interactions involving multiple entities [28].

Jamieson *et al.* used text-mining to recreate the HIV-1, Human Protein Interaction Database (HHPID) [29]. Protein NER was carried out by BANNER [30] while interactions in text were identified using 2 tools, the Turku event extraction system (TEES) [31] and EventMiner [32]. The NER and event extraction was applied to 3090 titles and abstracts and 49 full-text articles

achieving a precision, recall and *F*-score of 87.5%, 90.0% and 88.6%, respectively. The pipeline was able to completely replicate over 50% of the database. The team observed that the greatest obstacles to the automated extraction were grammatically-complex sentences and sentences containing poor grammar.

While the study largely mined text from article abstracts, the team considered the use of more full-text articles in the future. They identified significant challenges that full-text articles could introduce, however, such as distinguishing true reactions from hypothetical reactions in the Discussion and retrieving full-text articles in the first place (a challenge that I describe in depth in Section 5).

While new methods for extracting PPIs are regularly released [33, 34], attention is increasingly shifting towards more complex relationships, with a particular focus on biomolecular networks and pathways [35] such as protein–protein interaction networks [36, 37], signal transduction pathways [38, 39, 40], protein metabolism (synthesis, modification and degradation) [35], and regulatory networks [41, 42]. This protein/gene-centric focus has been enshrined in most of the competitive text-mining events (such as BioCreative [43, 44, 45] and BioNLP [35]). However, in spite of this new focus on networks and pathways, one of the most important subtopics — the construction and curation of metabolic pathways — has largely been ignored.

4 Metabolic interaction extraction

The only system that I am aware of that has an explicit focus on extracting metabolic pathway information from free text is the template-based EMPathIE [46], which is no longer under active development (R. Gaizauskas, personal communication). The aim of EMPathIE was to extract information about metabolic reactions together with relevant contextual information (including source organism and pathway name) from specific journals. When evaluated on a corpus of seven journal articles, EMPathIE achieved 23% recall and 43% precision [47].

Certain more generic systems may also be used for the same purpose, including the GeneWays system for "extracting, analyzing, visualizing and integrating molecular pathway data" [38], and the MedScan sentence parsing system [48], capable of extracting relationships between a range of biomedical entities including proteins and small molecules, and evaluated on a PPI extraction task by Daraselia *et al*. [49]. However, neither GeneWays nor MedScan

are freely available and I am not aware of any published evaluation of their performance with metabolic pathway data.

It is interesting to note that the creators of GeneWays, in that system's key publication, suggest signal-transduction pathways are an "easier target" for information extraction than metabolic pathways, and chose to evaluate its performance on the former rather than the latter [38]. Similarly Hoffmann *et al*. identify the extraction of metabolic information as a "special case" that has "specific problems" associated with it [50]. This perception may explain why relatively little attention has been paid to the task of extracting metabolic reactions from free text. The particular challenges that are characteristic of metabolic reaction extraction tasks include:

- Multiple entity types and entity mismatch. Whereas protein-protein interaction networks, protein metabolism and signal-transduction pathways concern the entity-type protein, metabolic reactions involve both enzymes and metabolites. Moreover, there is a mismatch between the entities that most taggers address (proteins/genes, small molecules) and the entities involved in metabolic pathways (principally enzymes and metabolites). Similar problems arise in the context of the extraction of protein-protein interactions owing to the fact that protein/gene taggers almost invariably fail to distinguish between proteins and genes. Only a subset of proteins are enzymes, and whereas the distinctive nomenclature associated with enzyme names may be beneficial to the extraction process (such as the suffix *-ase*), it has been argued that identifying the names of metabolites is more difficult than some other categories of chemical name [51].

- Ternary (and *n*-ary) relationships. Whereas the relationships in protein-protein interaction networks and signal-transduction pathways are typically binary (e.g. "protein A activates protein B"), metabolic relations are typically ternary (e.g. "enzyme C catalyzes the conversion of substrate D to product E"). Moreover, multiple substrates and/or products are commonplace, leading to further complexity. One consequence is that there is a greater potential for all the relevant entities in a metabolic reaction to be split over multiple sentences and for there to be a high incidence of anaphora usage.

One of the initial goals of the project was to address the question as to whether the extraction of metabolic reactions is, indeed, more difficult than the extraction of protein–protein interactions.

Although the fully-automated construction of networks and pathways from the literature may be the ultimate goal, a more practical focus for text mining systems in the immediate future is to provide assistance to database curators and model builders. Existing initiatives specifically designed to support database curation include PreBIND [52] and various tools [53] aligned with the task of curating FlyBase [54]. In this context, high recall is often deemed to be of paramount importance, although excessive numbers of false positives detract from the usability of such systems [55]. Existing initiatives designed to assist the curation of pathway and network databases include research that addresses the curation of Wnt signaling pathways [39] and an application designed to support the curation of chemical–gene–disease networks in the Comparative Toxicogenomics Database [56].

5 The automated retrieval of journal articles

The challenge of retrieving full-text articles has long held back biomedical text-mining. All article titles and abstracts can be obtained using the mature and stable E-Utils API provided by the NCBI [2]. As the API allows article records, containing the article abstract, to be retrieved in bulk and in a common format, early text-mining work in the biomedical community concentrated on the mining of these easily obtainable abstracts. While mining abstracts can return important data (as the significant findings of a paper will be repeated in the abstract), a great deal of potential useful data is only found in the full article. There has been a clear move towards developing tools using full-text articles with the BioCreative III tasks using corpora of full-text articles for the competition [45].

The move towards full-text article analysis is not simply a case of having more text to process for each document, however. Cohen *et al.* compared the structure and content of abstracts with the article bodies [57]. They found that article bodies contained longer sentences and significantly more parenthetical material which would presumably hinder information extraction. Jamieson *et al.* found reading grammatically complex sentences was one of the main obstacles in extracting interactions [29]. A number of text mining tools were tested on the abstracts and article bodies separately. The results for the majority of tools were better when reading abstracts as opposed to article bodies, which came as no surprise as it was the standard to develop and train tools on abstracts.

Unfortunately, publishers have been reluctant to allow their publications to be mined. While it is possible to scrape articles from the publishers' websites [58], typically the Robots Exclusion Standard of most publishers' websites disallows access to screen scraping tools (with the exception of search engine spiders, such as the Googlebot). While the rules set by the `robots.txt` file are purely advisory and rely on the cooperation of the spider, web administrators can block access if they wish[4].

PubMed Central (PMC), a repository of full-text articles from free-access journals, provides an API (as part of the NCBI E-Utils API) to obtain complete articles in a common machine-readable format from the Open-Access Subset, in addition to bulk downloading the subset over FTP. The number of articles available in PMC significantly increased following the Consolidated Appropriations Act of 2008 (H.R. 2764) being signed into US law which required all NIH funded research to be freely accessible and available through PMC within 12 months of publication. Other research funding bodies, such as the 24 members of the Europe PubMed Central funders group have since followed suit [59]. It has now become standard in text-mining research to utilise full-text articles from PMC alongside abstracts retrieved from PubMed.

As I have alluded to, PMC contains two subsets of articles: open-access articles, which can be downloaded in full using the API or the FTP server, and free-access articles, which can only be viewed in full on the website. Figure 2 shows the number of records published in PubMed in each year since 1990. The rate of publishing is increasing and 2014 will certainly be the first year with one million new records. The graph also shows the number of articles originally published each year that are held in the PMC Free-Access and Open-Access Subsets. At the time of writing, the PMC Open-Access Subset contains approximately 750 000 articles compared to over 24 million article records held in PubMed. While the rate of open-access publishing is increasing, it is not undergoing the same explosive growth that publishing in general is experiencing and is falling further behind. Looking back, the number of open-access articles available that were originally published prior to 2005 is almost insignificant compared to the total number of published articles.

While the open-access model is useful for text-mining research, the non-open-access model is unlikely to disappear. Attitudes and laws regarding the text-mining of research are slowly changing, however. On June 1st 2014 the UK Government brought into force the UK text and

[4]While investigating the potential use of screen scraping I discovered that while most publishers allowed the relatively small scale scraping operation I employed, not all publishers were so lenient.

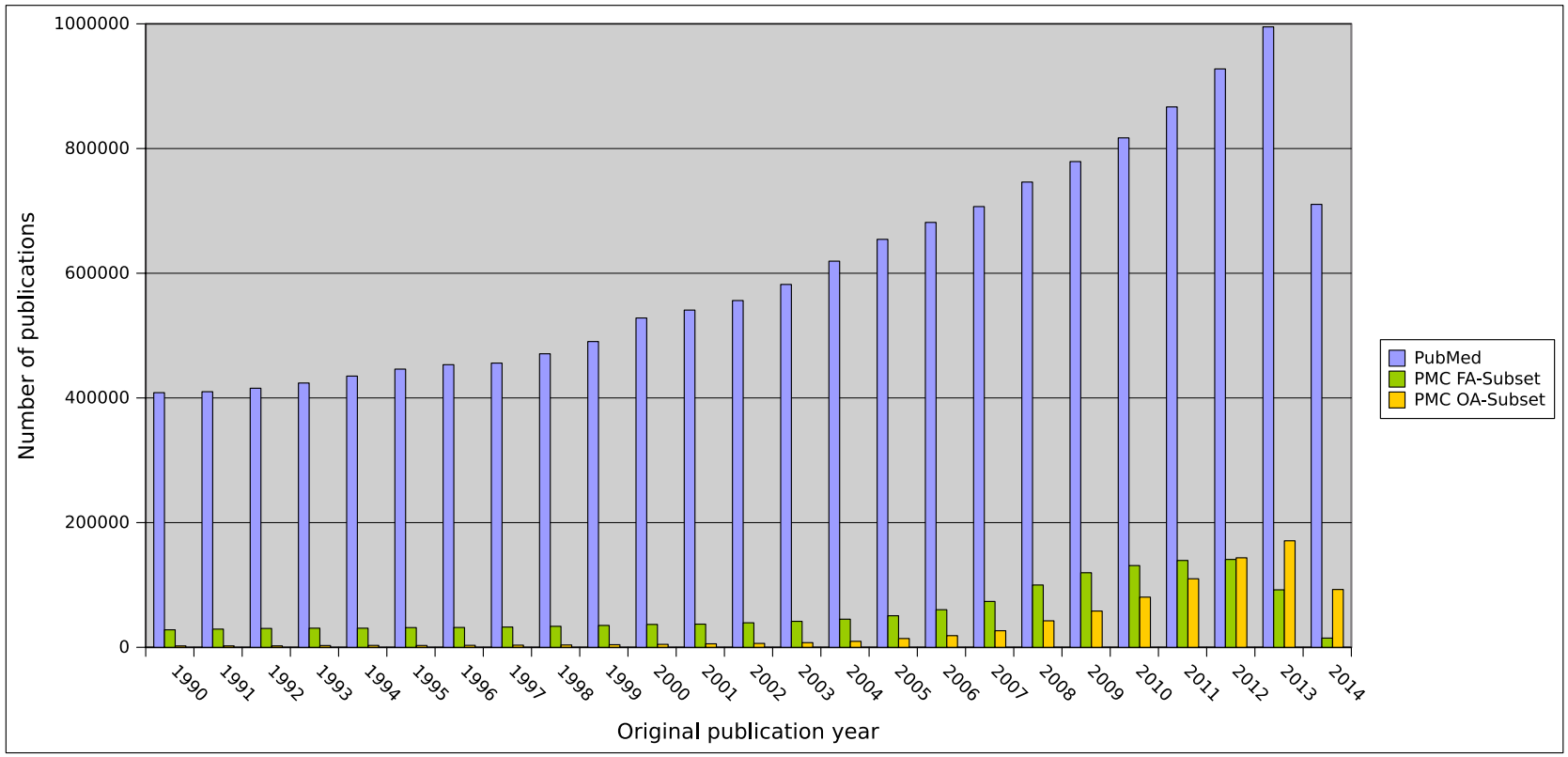

Figure 2: A graph showing the number of records added to PubMed and the number of articles in the PMC Free-Access and Open-Access Subsets originally published in each year since 1990 (including 2014 up to 30th July).

data mining exception, which allows researchers with lawful access to articles to copy them without explicit permission for the purposes of non-commercial text and data mining [60]. In response to this change some publishers, such as Elsevier, now provide an API to access their journals [61] and the CrossRef service has released a general API so that separate code does not need to be written for every publisher with a retrieval API [62].

These developments have not been met with open arms by the text-mining community, however, with researchers unhappy with the terms of use of the Elsevier API which they view as being overly restrictive [5]. For instance, the API is restricted to the retrieval of text, while figures remain inaccessible and the API can only be used for non-commercial uses. The most significant limitation of the API, however, is the explicit prevention of the automated crawling of content — rendering it practically useless for text-mining. The UK copyright exception forces no requirements on publishers on the access to the content.

In recent years the Associtation of European Research Libraries (or Ligue des Bibliothèques Européennes de Recherche — LIBER) have lobbied for changes to copyright legislation to allow the use of text and data-mining methods to extract data from content that researchers have access to. LIBER, in response to the terms of the Elsevier API [63], stated their belief that the right to read is the right to mine and that the introduction of licenses in addition to a subscription to a journal, such as those required by the Elsevier API, is "unscalable and resource intensive" and can only limit scientific progress.

The perspective of non-open-access publishers was described in the European Publishers' Council ("a high level group of Chairmen and CEOs of leading European media corporations") Copyright Vision 2014 [64], where the demands for automated access to scientific journal articles for the purposes of data-mining were viewed as follows (p. 48):

> In our view this is "a snare and a delusion" perpetrated by those intent on gaining free access to the widest possible body of copyright works in the name of research, going way beyond scientific journals, to works of all published authors, as well as Europe's news media and entertainment.

While it is difficult to get behind this wording[6], publishers are wary, perhaps justifiably, of the

[5] See #*ElsevierGate* on Twitter for the unfiltered community response to the terms of the Elsevier API.

[6] Wording which seems to suggest that the push for automated access to scientific journals is a conspiracy by people who want to download music and films for free. LIBER, unsurprisingly, responded with an open letter to the European Commission [65] in light of the European Union's upcoming copyright review.

possibility of their whole catalogue of articles being made available on the Internet by a lone unscrupulous researcher.

The release of the Elsevier API and the general CrossRef API unfortunately came too late in the project to analyse their worth. The reaction within the community has shown that this is an issue that promises to remain for some time yet, so the publicly available tool produced by this project will only allow the automated retrieval of abstracts and open-access articles. In Section 18.2, however, I show the stark difference in results between mining all available full-text articles and mining just abstracts and open-access articles.

6 LiMPET — A metabolic pathway extraction tool

There are different ways of positioning the system that I aim to create. Enabling the fully autonomous extraction and construction of pathways from the literature would be a lofty goal and perhaps too ambitious, at least in the short term. A more realistic aim would be to create a tool capable of aiding database curation — a task explored in BioCreative III in the development of tools capable of selecting articles suitable for the manual curation of protein-protein interactions [27]. Such tools exist in other fields. For instance, curators for FlyBase, the primary database of molecular data for the *Drosophilidae* family, utilises a GATE pipeline which has an NER module to mark up gene names and is then able to link these to other entities in the text [66]. Similarly the BRENDA companion databases, FRENDA and AMENDA, utilise text-mining methods to extract enzyme information from PubMed abstracts [5].

The biomedical literature is littered with well-performing text-mining tools that are not publicly available — a fact investigated further in Part II on currently available methods. While perhaps acceptable in the context of a competition, such as BioCreative, where the tools are often crude prototypes which have not undergone the sort of testing required to create a reliable, user-friendly tool, the open availability of both the final tool and the source code is necessary to advance knowledge in the field — particularly as ours is the first published attempt at metabolic reaction extraction since EMPathIE [46].

The ease of use of the tool is similarly important. Kabiljo *et al*. assessed the usability of a number of protein-protein extraction tools and found their use fraught with difficulties [1]. Releasing a difficult to use tool can be almost as bad as not releasing the tool at all. While I

have not developed a graphical user interface for my tool, I have endeavoured to make the tool easy to install and use on its own, in addition to collaborating with other groups to allow integration with other tools.

In this thesis I describe the development of LiMPET, the Literature Metabolic Pathway Extraction Tool. LiMPET is designed to be a relatively simple first attempt at solving the problem of metabolic pathway extraction; a baseline on which more advanced tools can be built on and compared to. In Part II I describe the third party software that was utilised and in Part III I describe the novel metabolic reaction extraction algorithm that forms the core of LiMPET. In Part IV I describe the use of LiMPET to recreate experimentally verified pathways in EcoCyc showingthat LiMPET, with its combination of a pattern based method and global association analysis, performs strongly in its own right. In Part V the use of LiMPET to corroborate predicted metabolic pathways in BioCyc is shown, showing that metabolic pathway extraction is a tractable problem that deserves more attention from the biomedical text-mining community.

Part II

An overview of text-mining methods

In this chapter I will describe the existing methods and tools that were utilised in the development of LiMPET, while the text-mining method developed during this project will be described in following chapters.

7 Approaches to text-mining

The text-mining projects I described in Part I and the libraries used in this project use a variety of text-mining methods.

The simplest text-mining methodology is dictionary-based. Dictionary-based methods utilise a predefined collection of terms which is used to find matches in the text of interest — making them particularly amenable to named entity recognition (NER) tasks. Dictionary-based methods have a very simple methodology and, if a suitable dictionary is already available, are very easy to implement. They cannot take context into account, however. Consider an NER system for the identification of given names in text. Using just a dictionary of common names it would be impossible to determine whether the term *Jack* was a person's name or the common English noun or verb. Dictionary-based methods are also unable to recognise novel terms that are not in the dictionary.

Heuristic (or pattern-based) methods rely on manually defined rules and are relevant to both named entity recognition and relationship extraction. Consider the development of an NER system for enzyme names. One could hypothesis that enzyme names could be recognised by identifying words with the suffix *-ase*. Such a heuristic method would be able to recognise novel enzyme names and it would also be possible to implement rules that take the context of the word into account. Kabiljo *et al.* developed a heuristic method for the extraction of protein-protein interactions which looked for predefined "interaction verbs" between two protein names [1]. As described in Section 3, however, human language is constantly evolving and rules are rarely defined clearly. Any rule implemented is likely to have significant exceptions that must be considered (e.g. not all enzymes have the suffix *-ase* and not all words with the suffix *-ase* are enzymes). Systems that deal with particularly complex language may require

input from a linguist in the development of rules.

Machine learning is an umbrella term for a large number of varied statistical methods which all have the ability to learn from examples to create a general model that describes the data. In the field of text-mining, machine learning methods learn from corpora of text (of the type that the method is intended to be used to analyse) with all entities and relationships of interest annotated. While machine learning methods are typically able to outperform the simpler methods I have described, the need for large quantities of annotated text to train new models can hinder their implementation in fields where a training corpus is not readily available.

In the areas of protein NER and PPI extraction, conditional random fields (CRFs) [67] have become one of the most popular methods (see Sections 3.1 and 9.3.1). Unlike methods such as naive Bayes classifiers, CRFs are able to take context and sentence structure into account, and unlike hidden Markov models, have been shown to not suffer from label bias (where states with fewer possible transitions are favoured over those with a greater number of possible transitions as the node with fewer possible transitions will typically have higher probabilities [67]).

8 Performance assessment

The quantitative assessment of text-mining systems tends to involve the use of corpora — collections of documents with entities and relationships manually annotated. The system being assessed is run on the text within a corpus and the results compared to the marked-up elements using the following measures:

Precision The proportion of extracted instances that are correct extractions.

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall The proportion of relevant instances that are correctly extracted.

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

F_1-score (Often abbreviated to **F-score**) An overall measure of accuracy — the harmonic mean of precision and recall.

$$F_1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall}$$

These measures rely on the availability of a suitable corpus, however. The creation of a corpus is a very time consuming task. There are many reasons why annotating articles can take a significant amount of time, but one practical concern is the tools used to write the annotations. Typically, annotated documents are written in XML (a computer markup language). Manually writing XML can be a labourious, error-strewn process for even an expert in the language. Tools have been developed, however, to allow the annotation of text with no deeper knowledge of the underlying format and allow the curator to concentrate on the annotations themselves. PubTator is a web based tool to assist the curation of PubMed abstracts [68]. Abstracts are automatically marked up with text-mining tools, but annotations can be easily changed and added. While there are plans to expand its use to full-text articles, this will likely be limited to open-access PubMed Central articles.

Moreover, accurately annotating text can require specialist domain knowledge where the subject matter is complex or the text seems ambiguous to an untrained eye. In the development of the BioCreative I training and test data, it took on average one week for a single curator to curate 250 abstracts [69]. Additionally the accuracy of annotations was checked by cross-referencing a subset of abstracts that were annotated by multiple curators. In the case of abstracts concerning mouse genes, only 69% of annotations were agreed by three different curators.

There is precedence, however, for evaluating text-mining tools in the absence of a suitable corpus. Yuryev *et al*. developed a method for the construction of "biological association pathways" from the redundant networks extracted using the tool MedScan [40]. To assess their method, the tool was used to reconstruct manually constructed pathways based on review articles.

Rodríguez-Penagos *et al*. developed a tool for the extraction of regulatory networks from text [42]. As a substitute for an annotated corpus, sets of abstracts and full-text articles that were potentially relevant to *Escherichia coli* K-12 were collected from references in relevant databases (RegulonDB and EcoCyc) and by using carefully-crafted search strategies. It is not possible to automatically calculate a precise recall using this strategy as it is impossible to determine if the extraction method has missed any extractions without manual analysis. The group instead compared the extracted relationships to those in RegulonDB, finding that the extracted relationships covered 45% of the database. As an estimate of precision, a biologist examined a

random sample of 96 interactions which did not map to known interactions in RegulonDB and found 81 had a basically correct semantic interpretation of their source sentences — giving a precision of 84%.

As there is no available metabolic reaction corpus, an evaluation strategy similar to those developed by Yuryev *et al*. [40] and Rodríguez-Penagos *et al*. [42] was used in the evaluation of LiMPET's core text-mining algorithm (see Section 13.2).

8.1 Assessing ranked extractions

When extracting reactions there isn't a clear definition of what reactions are relevant to the user — there is a certain level of subjectivity. The tool would not be particularly useful, however, if a simple list of extracted reactions (which may be considerable in length) are returned to the user. It is, therefore, desirable for extractions to be ranked according to their potential relevance to the query. The traditional measures of recall, precision and F-score do not take ranking into account, however — extractions are considered to be either correct or incorrect.

Swets [70] investigated a number of retrieval performance measures and determined that a desirable method would have the following properties:

1. It would only be concerned with the ability of the system to differentiate relevant and irrelevant items and would not be affected by other factors, such as efficiency.

2. It would be independent of any scoring threshold (whether from the user or a characteristic of the retrieval system) and would measure the system's total output.

3. It would be a single number, as apposed to a pair of numbers or a curve, so as to facilitate easy comparison between methods.

4. It would have absolute significance as a measure of a single method and allow comparisons of different methods.

Swets outlined these requirements with general retrieval tasks in mind where retrieval scores tend not to be shown. In bioinformatics, however, retrieval scores are typically provided to the user and, moreover, a scoring threshold is typically applied to the list (which can be altered by the user) and items not meeting the threshold are ignored. This led to Wilbur [71] modifying condition 2 to bring it further in line with the use of retrieval systems in bioinformatics:

It should be characterized by a [user] threshold, but should reflect the quality of retrieval at every rank down to that threshold.

Carroll *et al*. [72] introduced "The Principal of Fidelity", which states that: "a retrieval measure should faithfully reflect the actual usage of the retrieval list". Considering this, Carroll *et al*. introduced the following three conditions to follow on from the Swets' and Wilbur's conditions:

5. It should be robust against results representing a small proportion of possible user queries.
6. When two disjoint sets of queries are considered, its value for the union of the two sets should lie between its values for the two sets of queries.
7. It should reflect the choice of threshold; in particular, it should eventually decrease as the threshold increases to include the entire retrieval list.

8.1.1 ROC analysis

ROC analysis was a popular method in clinical applications to evaluate the performance of diagnostic tests in diagnosing specific medical conditions, when Gribskov & Robinson [73] recognised the potential of the analysis to evaluate the performance of sequence annotation methods. The initial steps of ROC analysis are the assignment of each datapoint as positive or negative (relevant or irrelevant in the case of information extraction) and the subsequent construction of a ROC curve where each point indicates the fraction of positives equal to or greater than a specific score on the x-axis and the fraction of negatives equal or greater than the same score on the y-axis. The following steps are employed to calculate the ROC curve:

1. The retrieval system produces a ranked list of extractions from most to least relevant.
2. Manually mark each item as relevant or irrelevant.
3. At each ranking calculate i, the number of relevant items at or preceding the ranking.
4. Number each irrelevant item in the list 1, 2, ..., f, ..., F according to its ranking.
5. The ROC curve plots i/I (where i=the number of preceding relevant items and I=the total number of relevant items), the fraction of relevant items preceding the f^{th} irrelevant record, against f/F, the fraction of irrelevant items seen.

1	2	3	4	5	5
Ranking	Relevance	Cumulative relevant items (i)	Cumulative irrelevant items (f)	i/I	f/F
1	1	1	0	0.2	0
2	1	2	0	0.4	0
3	0	2	1	0.4	0.1
4	1	3	1	0.6	0.1
5	0	3	2	0.6	0.2
6	0	3	3	0.6	0.3
7	1	3	3	0.8	0.3
8	0	4	4	0.8	0.4
9	0	4	5	0.8	0.5
10	0	4	6	0.8	0.6
11	0	4	7	0.8	0.7
12	0	4	8	0.8	0.8
13	0	4	9	0.8	0.9
14	1	5	9	1.0	0.9
15	0	5 (I)	10 (F)	1.0	1.0

Table 2: Example ranked data showing the calculation of the corresponding ROC curve. Numbers in the top row correspond to the step employed to calculate the data in the column (see page 36).

Table 2 shows example ranked data that is used to create the ROC curve in Figure 3 using these steps.

A data set with good discrimination between relevant and irrelevant items will produce a curve that lies in the upper left of the graph area (as can be seen by the blue curve corresponding to the example data in Figure 3). A data set with no discrimination, however, will produce a curve similar to the orange curve. The ROC score is the probability that a random relevant record is ranked above a random irrelevant record and is equal to the area under the graph. The example data produces a ROC score of 0.74, while the curve showing no discrimination has a ROC score of 0.50.

In a typical biological application, however, the number of negatives in the whole dataset will vastly outnumber the number of positives. This means that even a merely adequate ranking of items will produce a ROC score close to 1. For this reason it is typically necessary to

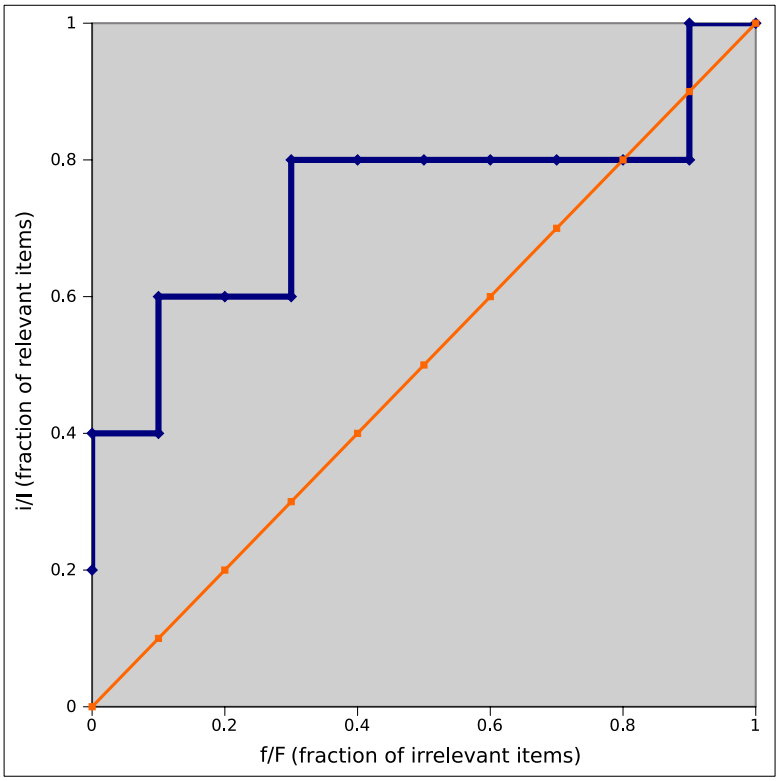

Figure 3: An example ROC curve created from the data in Table 2. The blue line corresponds to the example ranked data, while the orange line shows data for a notional method that is unable to discriminate relevant and irrelevant items.

---

calculate a ROC_n curve, the ROC curve truncated after n irrelevant records, and the ROC_n score, the area beneath this curve divided by n/F. The value of n can be altered depending on the use case, but a threshold of n=50 is common practice [73]. To combine the results from different runs of a particular algorithm a pooled ROC curve and pooled ROC_n score can be produced by merging the ranked retrieval lists from the separate queries into one list.

ROC analysis has been used previously for evaluating the performance of text-mining systems. Frijters *et al*. [74] developed CoPub Discovery, a co-occurrence method for the discovery of drug, gene and disease connections in text. In the absence of a suitable corpus, the tool was evaluated on its ability to identify true positive and false positive hidden relationships in a partitioned set of Medline abstracts. Separate ROC curves and scores were produced for different types of relationships (e.g. gene-disease, drug-disease) and were not directly compared, while different results for different scoring criteria for each relationship type, were compared.

8.1.2 Precision-recall (PR) curves

The use of ROC analysis for evaluating information retrieval has been criticised, however, with the use of precision-recall curves, which simply plot the precision and recall at each relevant item in the retrieval list, being put forward as a suitable replacement. PR curves are notably more suitable when the number of irrelevant items is far larger (typically by orders of magnitude) than the number of relevant items — the typical situation in information retrieval. For such data sets, ROC analysis requires the calculation of a ROC_n curve, as described previously, whereas the PR curve will be the same regardless of the number of irrelevant items found at the bottom of the ranking. The use of ROC_n curves can be problematic when comparing different queries and methods that produce rankings where different thresholds would be suitable.

The use of pooled ROC curves, and the resultant pooled ROC score, in information retrieval has been similarly criticised. Sierk and Pearson [75], in an evaluation of protein structure comparison methods, found that pooled ROC curves could be unduly distorted by the poor performance of a small number of queries and instead examined the performance of the methods with individual queries. Carroll *et al*. showed an example where the pooled ROC curve of two queries is lower than both of the individual ROC curves [72].

Table 3 shows the same ranked data as before with the precision and recall calculated at each relevant item. The average precision is the mean of the precision at each relevant item in the retrieval list. Figure 4 shows the resultant precision recall curve. A data set with no discrimination will achieve an average precision of about 0.5 regardless of recall, resulting in the orange curve as can be seen in Figure 4, whereas a data set where relevant items are clustered at the top of the ranking will produce a curve that is higher (for most of the curve, at least) and a higher average precision.

8.1.3 TAP-k

Average precision does not take into account a user threshold, however, which breaks Wilbur's modification to the second condition of a good retrieval measure [71] described previously ("It should be characterized by a [user] threshold, but should reflect the quality of retrieval at every rank down to that threshold."). In response to this Carroll *et al*. developed the measure Threshold Average Precision at a median of k errors per query (TAP-k) which is based on average precision, but reflects the user's tolerance for errors [72]. The TAP score at a particular

Ranking	Relevance	Precision	Recall
1	1	1.00	0.2
2	1	1.00	0.4
3	0		
4	1	0.75	0.6
5	0		
6	0		
7	1	0.57	0.8
8	0		
9	0		
10	0		
11	0		
12	0		
13	0		
14	1	0.36	1.0
15	0		
Average Precision (AP)		0.66	

Table 3: Example ranked data showing the calculation of the corresponding PR curve.

scoring threshold is defined as follows:

1. The retrieval system produces a ranked list of extractions from most to least relevant.
2. Manually mark each item as relevant or irrelevant.
3. Define a scoring threshold and assign the first item prior to the threshold as the 'sentinel'.
4. Calculate the precision at the sentinel and at each relevant item with a score greater than the sentinel.
5. Assign a precision of 0 to each relevant item scoring lower than the sentinel.
6. Calculate the mean of the precisions from steps 4 and 5 to calculate the TAP.

Table 4 shows the example data with the associated precisions at a number of thresholds.

To calculate the appropriate threshold for the results from a series of queries, the user is assumed to tolerate k errors per query (EPQ). The lowest threshold that produces a median EPQ

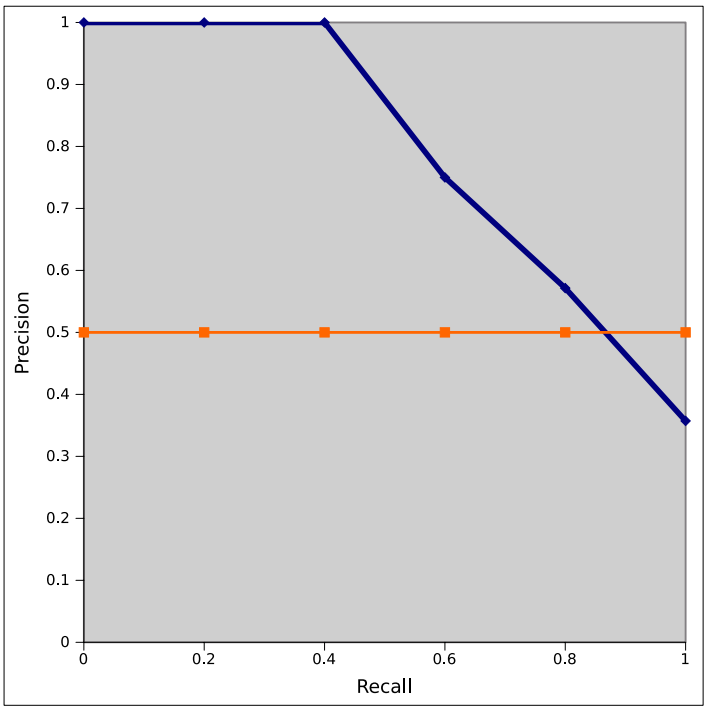

Figure 4: An example precision-recall curve created from the data in Table 3. The blue line corresponds to the example ranked data, while the orange line shows data where there is no distinction between relevant and irrelevant items.

---

of k across the queries is chosen as the threshold. While the value of k is arbitrary and may depend on the problem at hand, the values of 5, 10 and 20 were chosen for the gene normalisation task at BioCreative III [45]. While not directly comparable to the metabolic reaction domain, the TAP-k scores achieved by the competing tools provide the only possible comparison to the scores achieved by LiMPET.

As the development of LiMPET has progressed, the characteristics of results produced and the methods used to evaluate them have changed. Early in the development process LiMPET was evaluated by extracting reactions from a small selection of hand-picked relevant papers using a variant of the F-score (see Part III) while later in the project the tool was tasked with extracting reactions from a large number of automatically retrieved articles, of unknown relevance, where extractions were ranked and the TAP-k measure was used (see Part IV).

Ranking	Score	Relevance	Precisions with threshold		
			0.65	0.75	0.90
1	1.00	1	1.00	1.00	1.00
2	0.95	1	1.00	1.00	1.00
3	0.90	0			0.67
4	0.85	1	0.75	0.75	0
5	0.80	0			
6	0.75	0		0.50	
7	0.70	1	0.57	0	0
8	0.65	0	0.50		
9	0.60	0			
10	0.55	0			
11	0.50	0			
12	0.45	0			
13	0.40	0			
14	0.35	1	0	0	0
15	0.30	0			
		Average (TAP)	0.64	0.54	0.46

Table 4: Example ranked data showing the calculation of TAP scores at three different thresholds. Precisions in blue signify the assigned sentinel records at each scoring threshold. Precisions in green belong to relevant items with a score above the threshold and precisions in red belong to items below the threshold which are all given a precision of 0.

9 Third party tools

In order to effectively utilise the time available for the project, I attempted to incorporate available software into the tool whenever possible. I had the following requirements of any software to be incorporated:

- The software must be open-source with no restrictions on incorporating it into other tools.
- Ideally the software would be available as a Java library. While it is possible to incorporate command line programs into a Java program, this may hinder the easy running of the tool and may prevent it from running cross platform. The incorporation of web services should only be used sparingly as mining a large amount of text could cause too many requests to be sent to the service.

- Databases must be freely available and downloadable to embed into the tool. Network access to a database can only be allowed sparingly for reasons described in the previous point.

9.1 A text-mining framework

It would certainly be possible to program the tool from scratch, but there are many advantages to utilising a text-mining framework. The tool was intended to be open source and freely available for others to view and fork the code. Writing a program from scratch could hinder the ability of others to edit the code, particularly for a complex project such as this. Anybody who is familiar with the framework used to develop a program, however, should be able to understand the code much more quickly. Even if they are not familiar with the framework, the framework will provide its own documentation. In addition, frameworks are typically written in a modular fashion so that components can be reused between projects — further easing their incorporation into third-party software.

The first popular, open-source, general text-analytics framework was the General Architecture for Text Engineering (GATE) which began development in 1995 at the University of Sheffield and is still well maintained today [76]. GATE has grown into a very mature package which can be used as a standalone program or incorporated into software as a Java library. GATE also comes with many general text-mining modules (such as a tokeniser, sentence splitter and a part of speech tagger).

In 2009, the Unstructured Information Management Architecture (UIMA) was made an OASIS standard [77, 78]. UIMA, initially developed by IBM and open-sourced in 2006, is a general framework that, unlike GATE, does not contain any text-mining components, but rather describes a standard way of constructing an information extraction pipeline.

UIMA has been utilised in a wide variety of applications, such as the clinical Text Analysis and Knowledge Extraction System (cTAKES) [79] and, famously, the IBM Research computer Watson which won an episode of the US quiz show *Jeopardy!*. UIMA has also been a popular choice of framework for the mining of biomedical research literature with many groups releasing their code as UIMA components (or with available UIMA wrappers) — a number of which will be described in the following sections. UIMA also forms the core of the U-Compare system — a graphical tool that allows users with no programming experience to create their own

text-mining pipeline using UIMA components [80].

While GATE is a very capable text-mining framework, I chose to use the UIMA framework, because of its popularity in biomedical text-mining and the availability of components written by other groups (many of which are discussed in the following sections).

9.2 General text-mining tools

Libraries exist for most programming languages for carrying out basic natural language processing tasks such as sentence parsing and part-of-speech tagging. One toolkit was found to fit our criteria: Apache OpenNLP [81].

OpenNLP is a machine learning based Java library and, as such, requires extensive training to create a suitable probabilistic model. While a large number of models are provided alongside the library, they are all the result of training the library on general-use language (typically from newspaper articles). The language used in biomedical articles, however, is highly specialised [82]. Buyko *et al*. [83] showed that transferring OpenNLP components to the biomedical domain was as simple as retraining the tool using a biomedical corpus, however, and that a specially designed tool was not necessary — for the low level text-mining tasks that OpenNLP deals with, at least. OpenNLP was retrained separately on two corpora, GENIA [84] and PennBioIE [85], and the subsequent performance of five OpenNLP components was assessed. Each component performed well when trained with either corpus, with the sentence splitter, tokenizer and parts-of-speech tagger achieving accuracies of approximately 99% and the chunker and parser achieving average F-scores of 92% and 86%, respectively. The group have released the trained models [86], allowing OpenNLP to be used with no need of further training.

OpenNLP was incorporated into the previously mentioned Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) where a number of components were built on OpenNLP components trained on clinical data [79].

9.3 Named Entity Recognition (NER)

Named entity recognition, typically the first step taken in a text-mining operation, aims to find entities within text and assign each to a predefined category. In this project it was necessary to be able to recognise genes and proteins, small molecules and organism names. Fortunately

there has been extensive research into each of these areas.

9.3.1 Gene/protein NER

There is significant subjectivity in what gene and protein names consist of. While corresponding genes and proteins typically have distinct names, they are often used interchangeably in text. For instance, the gene *ADH1* may be referred to as *the alcohol dehydrogenase gene* or *the gene encoding alcohol dehydrogenase*. Similarly the protein *alcohol dehydrogenase* may be referred to as *the protein encoded by ADH1*. Due to these difficulties and the unimportance of the distinction in many situations, tools rarely attempt to make any distinction.

The first freely available, user-friendly tool to attempt to solve this problem was ABNER, originally released in 2004 [87]. The tool included a graphical user interface and a Java programming interface to allow its functionality to be incorporated into other programs. ABNER was able to recognise protein names with an F-score of 84.9% and became the benchmark system with which to compare future work with.

Gene/protein NER was the subject of one of the tasks of BioCreative I in 2004, bringing together 15 groups to attempt the challenge [43]. The task called for the recognition of genes and proteins in three different organisms: yeast, fly and mouse. While the tools performed very well on yeast names (producing a high F-score of 92%), fly and mouse names proved more difficult to recognise (producing high F-scores of 82% and 79%, respectively). It was found that fly names had a significant overlap with English words and mouse names were more complex than yeast names. Overall, the highest F-score achieved was 83.6%. This was not significantly different from ABNER's score of 83.7% on the same corpus [88].

While most of the tools used a variety of statistical methods (principally hidden Markov models and support vector machines), one tool, Text Detective, implemented a rule based approach [89]. The tool tokenises a document and uses a variety of rules and dictionaries to categorise each token. For instance, terms such as kinase, receptor and transporter are highly indicative of a gene/protein name and are typically accompanied by chemical terms (which can be recognised by a combined rules-based and dictionary approach) and other biological terms which are stored in a dictionary. The tool also attempts to recognise gene symbols (such as TNF and p53) by recognising them as non-word tokens and using the precomputed probabilities of words that are typically found in the vicinity of gene mentions. While Text Detective

performed well on yeast and mouse gene names, no attempt was made on fly names.

BioCreative II in 2006 also proposed a gene NER task [44]. The best performing tool produced an F-score of 87.2% — an increase over the best score at BioCreative I. The entries to the second competition showed a significant move towards the use of conditional random field models in their methods. Unfortunately none of the tools entered were made publicly available, so ABNER remained the easiest to use, freely available system.

Leaman and Gonzalez, recognising this lack of freely available tools, developed BANNER, an open-source gene NER tool based on conditional random fields [30]. BANNER achieved an F-score of 82.0% — coming between the 9th and 10th ranked entries of BioCreative II. However, this evaluation was run before the testing corpus had become available and, therefore, the training corpus of 15 000 sentences was split in two to create a testing corpus. ABNER achieved an F-score of 78.3% on the same test corpus. This performance was repeated by Kabiljo *et al*. [1] who found that BANNER outperformed ABNER on four different corpora. BANNER is available as a Java library and, since its first publication, a UIMA wrapper has been included.

BioCreative III proposed a gene normalisation task, of which the first step was gene NER [45]. As gene NER was no longer the focus of this competition, most entries utilised previously developed NER tools instead of building their own from the ground up (although most entries involved sort type of post-processing of results from the NER tools). The NER tools used included ABNER and LingPipe, as well as BANNER and GNAT (which implements BANNER in its own method) [90].

I adopted BANNER for this project due to its good performance and its ease of implementation in a UIMA pipeline.

9.3.2 Small molecule NER

Small molecule NER methods use either dictionary or machine learning methods, or a combination of the two. This is due to there being two significantly different ways of naming small molecules — using the IUPAC systematic approach or using vernacular names. For instance, the molecule with the systematic name *butane-1,4-diamine* has also been given the vernacular name *putrescine*. Furthermore, a combination of the two approaches can be used — *N-carbamoylputrescine*, for example.

A dictionary-based approach can perform better with vernacular names than a statistical

approach as there is no common theme in the vernacular naming of small molecules and there are existing large and well-maintained databases of molecule names (which are discussed later in this section). Systematic names, however, are ideal for machine learning methods as they consist of a limited range of building blocks, but with very variable grammar. Kolářik *et al*. discovered that systematic and semi-systematic names could be recognised by their morphological structure with higher accuracy than with dictionary methods [91]. Because of this variability in naming molecules, well-performing small molecule NER systems typically utilise both approaches.

Narayanaswamy *et al.* [92] describe a prototype system designed to extract systematic and semi-systematic chemical names. Their method was based on a set of manually developed rules, as opposed to machine learning, as their were no comprehensive annotated corpora of gene/protein or chemical names at that point in time. The method produced promising results (an F-score of 81.69%) on a small test corpus of 55 abstracts.

In 2006, Corbett and Murray-Rust officially released the first freely available chemical NER tool: OSCAR3 [93], followed by OSCAR4 in 2011 [94]. The first tool to compete against OSCAR4, ChemSpot, was released in 2012 [95]. The only comprehensive comparison of these tools is found in the ChemSpot paper where both tools were tested against the SCAI chemical corpus [91]. OSCAR4 achieved an F-score of 57.3% while ChemSpot achieved 68.1%. As this project began in 2010, I initially implemented OSCAR3 and later updated to OSCAR4. Despite ChemSpot's performance advantage I decided to keep OSCAR4 as it had continued OSCAR3's good performance within the algorithm described here and there were more pressing aspects of LiMPET requiring focus. OSCAR4 is available as a Java library with an easy to use API. It was necessary to write a UIMA wrapper, however.

In addition to identifying small molecules in the text, it is also necessary to link together separate mentions of the same entity (in order to determine separate mentions of the same reaction and to link reactions together to form pathways). Chemoinformaticians have long understood the need for a common chemical language to describe small molecules. SMILES (Simplified Molecular-Input Line-Entry System), developed at the United States Enviromental Protection Agency in the 1980s, allows molecules to be described using ASCII strings [96]. SMILES has certain shortcomings, however, which rendered the format unsuitable for comparing molecules in separate databases. Principally, it is possible to obtain different, but valid,

SMILES strings from the same molecule. For instance, consider the molecule *α-D-glucose* which has a different SMILES identifier in two databases:

- ChEBI: `OC[C@H]1O[C@H](O)[C@H](O)[C@@H](O)[C@@H]1O`
- PubChem: `C([C@@H]1[C@H]([C@@H]([C@H]([C@H](O1)O)O)O)O)O`

In 2006, however, a new identifier was proposed: InChI (IUPAC International Chemical Identifier) [97]. Similarly to SMILES, InChI allows molecules to be described in ASCII strings, but InChI can express more information and, most importantly for our use case, every chemical structure has a unique InChI. With its greater utility and its backing from IUPAC, InChI quickly surpassed SMILES as the standard chemical structure identifier used by all major chemical databases. While it has been shown that there are inconsistent InChI assignments within and between databases, these inconsistencies are due to quality control issues and not the InChI format [98].

Due to the strict regular grammar of the IUPAC specification, it is theoretically possible to calculate the composition and structure of a molecule from its systematic name, a task which can be carried out by the Open Parser for Systematic IUPAC Nonmenclature (OPSIN) [99], which is included in the OSCAR4 library. The language used in chemistry literature can be very different to biomedical literature, however, where vernacular or semi-systematic names are far more common than strict systematic names. While this is understandable for complex molecules with suitably complex systematic names, even simple molecules are commonly described using older vernacular names instead of their systematic counterparts. For instance, *acetic acid* is commonly found in biomedical literature, while its systematic name *ethanoic acid* is rarely used[7].

Often systematic names are built with the correct components, but in the incorrect order. The following passages show alternative IUPAC-like names for the molecule *butane-1,4-diamine:*

> Linear or star-shaped poly(glycerol metha crylate)s (PGOHMAs) modified with **1,4-butanediamine** and 1,2-ethanediamine (EDA) were synthesized and used as polycations. [100]
>
> Herein, we employed single-molecule imaging and spectroscopy techniques for

[7]When searching for *ethanoic* in PubMed, it is helpfully suggested that I have spelt *ethanolic* incorrectly.

the detection of photochemical reactions between **1,4-diaminobutane** (DAB) and CdSe/ZnS single QDs. [101]

The biological chemical space is much smaller than the synthetic chemical space. For this reason, authors will often miss parts of the systematic name which are deemed irrelevant in a biological context — this particularly applies to stereochemistry. For instance, if the term *glucose* is found in the biomedical literature, the reader can safely assume that the specific molecule referred to is *D-glucose* as this is the predominant enantiomer found in nature. Being rarely found in nature, *L-glucose* would be expected to include the stereochemistry in the name.

With these inconsistencies in systematic names and the use of vernacular names, a chemical dictionary is necessary to determine the InChI of extracted small molecules. There are many chemical databases with the largest, PubChem, currently holding over 50 million compounds [102]. The size of this database would make incorporation into an offline tool difficult and potentially cause incorrect mapping of metabolites to irrelevant molecules. For instance, over 15 million compounds in PubChem are obtained from patents, of which presumably only a small percentage are used as drugs.

Chemical Entities of Biological Interest (ChEBI) is a more focused database currently cataloguing close to 40 000 entities [103]. While PubChem extracts compounds from many sources irrelevant to metabolic pathways, ChEBI principally sources data from four databases: IntEnz [104], KEGG COMPOUND [105], PDBeChem [106] and ChEMBL [107] — all of which are potentially relevant to metabolic pathways. With the small size of the database allowing easy incorporation into an offline tool and each specific compound containing a list of known synonyms and an InChI, ChEBI was chosen as the chemical database to incorporate into LiMPET.

9.3.3 Organism NER

Recognising mentions of organism names is necessary to determine the context of any entities or interactions found in an article.

While it might be possible to create either a rule-based or statistical method-based system able to recognise a given Latin species name, it would be unnecessary. While PubChem catalogues over 50 million compounds, all of which have a variety of possible names, there are relatively few organisms (NCBI Taxonomy currently contains approximately 420 000) and organism naming tends to follow strict conventions in scientific literature. Typically organisms

are referred to with their full Latin binomial name (e.g. *Catharanthus roseus*) or with an abbreviated genus (e.g. *C. roseus*). While many organisms have been given vernacular names (e.g. *Madagascan periwinkle*) they are generally used sparingly in scientific literature and typically confined to an introduction of the organism. Although one exception to this rule is with higher model organisms (such as *human* and *mouse*) where vernacular names may be preferred over taxonomic names.

There are, however, a number of complications that need to be taken into consideration. While there is no study in the literature on the accuracy of the spelling of Latin organism names, a feasible hypothesis would be that misspellings in Latin names are more common than in the rest of the text due to most people's unfamiliarity with Latin. Also, while full Latin names are designed to unambiguously identify organisms, use of an abbreviated genus can make the name ambiguous. For instance the abbreviated name *D. virginiana* could refer to *Didelphis virginiana* (the North American opossum) or *Diospyros virginiana* (the American persimmon). To a person reading this name in an article, the distinction should be obvious due to the context of the article, but establishing context is significantly more difficult for a computer program.

Latin names are, unfortunately, not static. Prior to genome sequencing, taxonomic classification (on which Latin names are based) was carried out by observing the morphological differences between organisms. The ability to compare genome sequences caused many of these earlier classifications to be revised, however. For instance, in 2007 the whole of the genus *Dryandra* was transferred into the genus *Banksia*, causing the renaming of many organisms, such as *Dryandra nivea* to *Banksia nivea* [108]. It is, therefore, necessary for any dictionary method to include all names that an organism has been assigned as the user may be mining older text.

TaxonGrab is a rule-based tool which identifies all words not found in an English language dictionary and applies rules to determine if a Latin organism name is present [109]. TaxonGrab achieved a recall of 94% and a precision of 96% against the 5000 page Volume 1 of "The Birds of the Belgian Congo" by James Paul Chapin, containing over 8000 taxonomic names. "Find all taxon names" is a tool based on TaxonGrab which implements additional rules to achieve a higher performance (>99% recall and precision) on the same evaluation set [110]. Neither tool was tested on their ability to normalise organism mentions against a database, however, and the evaluation corpus is unlike a biomedical research article. In addition, due to the rule-based

approach neither tool has the ability to recognise vernacular names.

There has been more focus on dictionary based methods, however, for the reasons elaborated previously and due to the free availability of a comprehensive and well-maintained organism database in the form of the NCBI Taxonomy database. The use of the database also automatically solves one problem with the rule-based methods: the normalisation of organism names against a database.

LINNAEUS is principally a dictionary-based method which implements some heuristic rules [111]. A dictionary of organism name synonyms was created using the NCBI Taxonomy database and abbreviated names were generated for each entry. Additional synonyms were identified that occur frequently in the literature — such as *patient* referring to *Homo sapiens*.

The group recognised the issue of ambiguous abbreviations and acronyms which can map to several organisms or even non-species terms (for instance, *PCV* can refer to *Peanut Clump Virus* or *Packed Cell Volume*). When LINNAEUS encounters an ambiguous term it searches for one of the possible expanded terms within the same text. It can even identify novel acronyms when defined in the format: *species (acronym)*, where *acronym* is a sequence of capital letters, digits or hyphens.

The tool performed well on a manually annotated corpus of 100 full-text articles from the PMC Open-Access Subset with 94.3% recall and 97.1% precision. The BioCreative III gene normalisation task required the entries to determine the source organism of genes in order to link them to database entries [45]. LINNAEUS was the only publicly available organism NER tool at the time and was used by the vast majority of teams. While the teams did note some ambiguity in species names and taxonomy IDs, the performance of LINNAEUS was well regarded.

LINNAEUS has since gained competition in the form of OrganismTagger, a hybrid rule-based/machine learning system [112]. While OrganismTagger was not able to perform as strongly as LINNEAUS, the novel machine learning approach could lead to better results in future versions.

LINNAEUS is available as a Java library and provides a UIMA wrapper for easy integration into a UIMA pipeline. These points and its good performance made LINNAEUS the obvious choice for integration into LiMPET.

9.4 Network visualisation

There are different ways to provide metabolic pathway data. In certain cases a user may prefer a list of the individual metabolic reactions or in a computer-readable language, such as SBML. In most cases, however, users will want to see the metabolic pathway output as an image which will allow them to easily parse the information. Indeed, visually displaying metabolic pathways is the standard method used to present information by databases such as BioCyc and KEGG. In order for LiMPET to be useful to users it must be able to construct such diagrams itself or provide the information to third-party tools to do so.

There are a number of general network visualisation libraries available for Java such as the Java Universal Network/Graph Framework (JUNG) [113], Graphviz [114] and Gephi [115]. Any of these libraries would allow a network to be drawn and shown to the user without leaving the Java application. Implementing functionality more advanced than simply displaying a static image to the user (such as the ability to manually edit the network), however, would be a significantly time-consuming task. It is also unlikely that any researcher using the tool would be doing so in isolation. Rather they would have multiple networks that would need to be compared and merged. This functionality is best achieved with an external tool.

While there are many standalone general network drawing packages, Cytoscape, originally created at the Institute for Systems Biology in Seattle in 2002, has become the standard visualisation tool for biological networks. Cytoscape has a number of advantages over a general visualisation tool. For instance, Cytoscape has the ability to import networks stored in systems biology file types (such as SBML [116] and BioPAX [117]), and is able to directly import Gene Ontology terms. Cytoscape also has a mature plugin framework and, due to the tool's popularity in the bioinformatics community, has a wide range of third-party plugins that enable sophisticated network analysis. For these reasons LiMPET was designed such that extracted metabolic pathways would be output in format viewable by Cytoscape.

Part III

A metabolic reaction extraction algorithm

In Section 3.1 I described the protein-protein interaction extraction task at BioCreative II where two principal methodologies were used by the competing teams: local and global association analysis [44]. In the development of LiMPET I have taken a hybrid approach. In the following chapter I will describe the local association analysis; the development of a core text-mining algorithm for recognising metabolic reactions described in individual sentences.

The work described in this chapter has been published [118] and the text here is largely based on the published article.

10 A methodology for extracting metabolic reactions

Various approaches have been utilized for extracting relationships between biological entities described in free text, broadly ranging from simple methods based on the co-occurrence of terms to sophisticated natural language processing methods. Here I adopt an intermediate, rule-based and pattern-matching approach that combines lists of stemmed keywords with rules for rewarding and penalizing the occurrence of words depending on their location. This approach can be viewed as an elaboration of several existing algorithms designed to extract protein-protein interactions (PPIs).

Indeed, the starting point for the algorithm developed here was the simple benchmark for PPI extraction presented in [1], which looks for ordered triplets of the form "protein name / interaction keyword / protein name". The Co3 algorithm, available via the Whatizit suite of Web services [26], takes a similar approach, as does the algorithm devised by Ono *et al*. [119], but with the addition of simple parts-of-speech rules.

This kind of algorithm is easy to integrate with established NER tools. The algorithm builds on two state-of-the-art named-entity taggers: BANNER [30] for recognizing gene/protein names; and OSCAR3 [93, 120] (later updated to OSCAR4 [94]) for identifying the names of chemical entities.

However, one important difference in the algorithm developed here arises from the intrinsic complexity of the relationships that are sought for extraction. For instance, small molecule

entities must be further classified as substrates or products. The algorithm assigns different permutations of such entities within a given sentence, with each permutation scored separately, although there are rules to ensure that implausible permutations are ignored. Details are given in section 12.2.

In terms of performance, one might anticipate that the algorithm will give higher precision, but lower coverage, than machine learning methods. It is interesting to note that the simple algorithm used in [1] proved remarkably effective when evaluated against some well-regarded machine learning approaches. When other factors are taken into account, such as execution speed and ease of installation, simple algorithms of this type are worthy of serious attention.

11 A metabolic reaction extraction task

It is first important to define what a metabolic reaction is. Broadly speaking a metabolic reaction is any chemical process that occurs in living organisms to maintain life. Typically, however, the term is used to describe the conversion of a set of non-peptide molecules into a different set of non-peptide molecules. The vast majority of such reactions in a living cell are catalysed by an enzyme. These reactions can occur in distinct pathways that serve many purposes — for example, the break down of large molecules to obtain the energy required to power the cell or the construction of large molecules, such as lipids, to store energy.

In reality, however, metabolic pathways are rarely distinct from one another, with molecules typically being involved in multiple pathways. In Section 2 I described two databases, BioCyc and KEGG, which take very different views on the organisation of metabolic pathways within a cell. In the development of LiMPET I have focused on the extraction of discrete pathways like those described in BioCyc, but due to the interconnected nature of metabolic pathways neighboring pathways are typically extracted as well resulting in raw extracted pathways that are more similar to KEGG pathways. In addition to the core extraction algorithm of LiMPET I have also developed functionality aimed at discerning the relevant information that is extracted by the tool from the irrelevant.

While the vast majority of chemical reactions are reversible, the use of enzyme catalysts result in most metabolic reactions having a clear direction. The initial molecules are referred to as substrates and the molecules that they are converted into are referred to as products.

The substrates and products are the most basic information required to describe a metabolic reaction, but the enzyme may also be included as well as descriptors of the type of reaction.

Not all metabolites in a reaction are equally important from an academic perspective, however. For instance, many varied metabolic reactions involve the breakdown of ATP to ADP and phosphate in addition to other metabolites. Therefore, when describing such a reaction in text, typically authors will not include these so-called *side* metabolites. While this does limit the scope of LiMPET (it cannot be expected to fully extract reactions which are not fully described in the text), the extraction of the *primary* metabolites remains a worthwhile task as it is these metabolites that define the purpose of the pathway.

LiMPET first uses the previously described NER tools, BANNER and OSCAR, to mark-up protein and small molecule entities in a document. The core algorithm described next attempts to determine which small molecule entities are substrates and which are products based on their ordering, position in the sentence and the presence of key words.

12 The algorithm

Given text in which the names of putative proteins and small molecules have been tagged, the algorithm proceeds in three key stages: a sentence selection phase; an entity assignment phase; and an assignment scoring phase.

12.1 Sentence selection

The algorithm begins by selecting sentences containing at least two small molecules. The working assumption is that sentences of interest will contain the names of both a substrate and a product, but not necessarily the name of an enzyme; it is sometimes possible to correctly identify substrate and product even when the name of an enzyme is not found (e.g. when it is mentioned in a separate sentence). In our training corpus (described in Section 13.1), 30% of the sentences that describe a metabolic reaction (and selected without reference to whether the name of an enzyme is present or not) do not contain the name of an enzyme.

12.2 Entity assignment

Given a selected sentence, potential orderings of putative enzyme(s), substrate(s) and product(s) occurring within the sentence are then considered in turn, or — in the absence of a putative enzyme name — orderings of substrate(s) and product(s). Manual analysis of text determined orderings that were highly unlikely to occur in practice. Specifically, it was determined that a particular entity type cannot be surrounded by entities of a different type (e.g. the order substrate - product - substrate). Such orderings were disregarded. The possibility that the reaction has multiple substrates and/or products is taken into account during this scoring phase. Consider, for example, the sentence:

> L-Arabinose isomerase catalyzes the conversion of L-arabinose to L-ribulose, the first step in the utilization of n-arabinose by *Escherichia coli B/r*. [121]

Here BANNER tags *L-Arabinose isomerase* as a putative protein, and OSCAR tags *L-arabinose*, *L- ribulose* and *n-arabinose* as putative small molecules. Ten different ways that the entities enzyme, substrate and product may be assigned to the tagged names are deemed suitable for consideration during the scoring phase. These assignments are given in Table 5.

12.3 Assignment scoring

The extraction algorithm described here is based on the baseline protein-protein interaction (PPI) extraction algorithm developed by Kabiljo *et al.* [1]. As the PPI extraction algorithm was concerned with extracting binary interactions by recognising a single keyword surrounded by protein entities, scoring was not necessary. Due to the relative complexity of metabolic reactions, however, with multiple entity types, multiple keyword types and different sentence structures it is necessary to score each component separately. Appropriate locations for each type of keyword given an ordering of enzyme and metabolite entities are defined in Appendix IV.

Given a sentence to which the entities substrate, product and (optionally) enzyme have been assigned, each assignment is then awarded a separate score based on the following criteria:

- Each entity within the assignment (either enzymes or metabolites) is awarded a positive score (+0.3 per entity).

L-Arabinose isomerase	L-arabinose	L-ribulose	n-arabinose
E	S	P	P
E	S	S	P
E	P	S	S
E	P	P	S
E	S	P	
E	S		P
E		S	P
E	P	S	
E	P		S
E		P	S

Table 5: Assignments of the entities enzyme (E), substrate (S) and product (P) for a sample sentence. The ten assignments of E, S and P for the sentence "L-Arabinose isomerase catalyzes the conversion of L-arabinose to L-ribulose, the first step in the utilization of n-arabinose by *Escherichia coli B/r*". Given that L-Arabinose isomerase is the only tagged protein, it is deemed to be the enzyme in all cases, whereas different numbers and orderings of substrates and products are possible, given the presence of three tagged small molecules (*L-arabinose*, *L-ribulose* and *n-arabinose*). Note that other potential orderings (namely E-P-S-P and E-S-P-S) are not considered, as they are deemed highly unlikely to occur in practice.

- Each word occurring between the first and last assigned substrate and product —- the entities *L-arabinose* and *n-arabinose* in the exemplar sentence above —- and that does not belong to the name of any additionally-assigned entities — *L-ribulose* in this exemplar sentence — incurs a small penalty (-0.1 points per word).

- If a keyword is found at an appropriate location relative to one or more entities (appropriate locations for reaction and production words are shown in Appendix IV), the assignment is awarded a positive score (+2 points per keyword).

- If a keyword is found in an inappropriate location, a penalty (of -1 point) is incurred.

- A bonus (of +2 points) is awarded when both a reaction and production keyword are found, provided they are in appropriate locations.

Keywords fall into the following categories: reaction word stems (e.g. *add*, *conver*, *hydrolys*, *dimeris*, *from*); production word stems (e.g. *form*, *give*, *produc*, *synthesi*, *to*); variants of the verb *catalyse*; and the coordinating conjunction *and* (all reaction and production word stems are listed in Appendix IV. Stemming was performed using a Java implementation [122] of the stan-

dard Porter stemming algorithm [123]. Scoring locations include: between an assigned enzyme and substrate for reaction keywords; between a substrate and product for reaction keywords, for production keywords and for the prepositions *to* and *into*; and between the last two assigned products/substrates for the word *and*. An example of an inappropriate location for a production keyword is before an assigned substrate.

As an example, here is the scoring for the exemplar sentence given the following partially incorrect assignments:

Enzyme = *L-Arabinose isomerase*

Substrate = *L-arabinose*

Products = *L-ribulose* and *n-arabinose*

- 4 entity assignments made: +1.2 points.
- Reaction keyword *conversion* found between the enzyme and substrates: +2 points.
- Production keyword *to* found between substrate and products: +2 points.
- Both a reaction word and a production word have been found: +2 points.
- Word *catalyzes* found: +2 points.
- Penalty for 8 words between first and last metabolite entities: -0.8 points.

This gives a total score of +8.4 points.

Consider the following correct assignments:

Enzyme = *L-Arabinose isomerase*

Substrate = *L-arabinose*

Products = *L-ribulose*

- 3 entity assignments made: +0.9 points.
- Reaction keyword *conversion* found between the enzyme and substrates: +2 points.
- Preposition *to* found between substrate and products: +2 points.
- Both a reaction word and a production word have been found: +2 points.
- Word *catalyzes* found: +2 points.
- Penalty for 1 word between first and last metabolite entities: -0.1 points.

This gives a total score of +8.8 points — higher than the score achieved by the incorrect assignments previously.

The list of assignments is ranked by score. Assignments below a threshold of 3.6 are removed from consideration. Pairwise comparisons of remaining assignments are made — if two assignments have at least one metabolite assignment in common, the lowest scoring assignment is removed (this allows multiple reactions to be extracted from a single sentence if they contain different metabolites).

The keyword lists and weightings used in the algorithm were chosen as follows:

- The reaction keyword list was assembled manually with specific reference to the nomenclature used in the Enzyme Commission (E.C.) classification [124].
  For instance, the E.C. classification *malate dehydrogenase (1.1.1.37)* leads to the stem *dehydrogenat* (matching words such as *dehydrogenates* and *dehydrogenation*).

- The production keyword list, together with the set of prepositions, conjunctions and addition reaction keywords, were assembled manually from an examination of the literature, from our own knowledge of the field, and using a thesaurus.

- The weightings (bonuses and penalties) used for each component when generating a score for a given assignment were derived from a small training corpus described in Section 13.1.

It is worth noting that an attempt to automatically compile a keyword list from verbs found between a tagged protein entity and a tagged small molecule in the GENIA corpus (a process analogous to that carried out by Kabiljo *et al*. in the context of PPI extraction [1]) proved insufficiently discriminatory to be useful, as the false positive rate was too high.

13 Training and evaluation

When considering how to evaluate this system, we found that existing corpora — even those with many sentences that contain the names of at least two small molecules, for example GENIA [125] and the metabolite corpus developed by Nobata *et al.* [51] — do not contain significant amounts of metabolic information relevant to the chosen target. Given that we perceive support for metabolic pathway curation as the ultimate goal of our research, we chose to assess

how many reactions belonging to a given metabolic pathway our system is able to extract from papers known to be relevant to that pathway. To this end, three contrasting pathways were chosen from the EcoCyc database [9] — a tier one database within BioCyc containing manually curated *Escherichia coli* metabolic pathways.

This approach to evaluation differs in two additional respects from the protein-centric BioNLP shared task on complex relationship extraction [35]: rather than abstracts alone, additional sections of full text articles were used; and entities were not pre-annotated. For the shared task, gold-standard annotations of protein names were provided from the outset. It was argued at the time that this would not have a major impact on the results, whilst it was acknowledged that it detracted somewhat from the task's "realism" [35]. However, an analysis by Kabiljo and coworkers demonstrated that the use of putative entity names that have been predicted using entity taggers (with an associated error rate of around 15%) instead of true, gold-standard entity names (extracted manually from the literature) can have a surprisingly large impact on relationship extraction scores, with "a fall of around 20 percentage points [in F-score] being commonplace" [1].

Although this approach to evaluation has been previously adopted elsewhere (see, for example, Rodríguez-Penagos *et al*. [42]), I acknowledge that it is "unrealistic" in that all papers are known to be relevant in advance. Although this is an important caveat, I believe the identification of relevant papers (e.g. with respect to the species of interest) is the task of a separate information retrieval component and that the evaluation of this system's ability to extract metabolic reactions is highly informative.

13.1 Training corpus

A small training corpus was used to set the weighting for the various scoring rules described in the previous section. This corpus consists of sentences containing the names of at least two small molecules selected manually from the literature referenced in the MetaCyc database [9] for various metabolic pathways, but excluding the specific pathways used subsequently for evaluation; 100 sentences were manually selected that describe at least one reaction each (with at least one named substrate and one named product), together with 100 sentences containing the names of multiple small molecules, but that do not describe a specific reaction. It is important to note that these were the only criteria used to select sentences from the set of referenced

papers. No attempt was made to exclude "difficult" sentences, hence the corpus contains the following complex sentence with multiple reactions:

> ZEP catalyses the epoxidation of zeaXanthin to produce epoxycarotenoid; NCED catalyses the cleavage reaction of epoxycarotenoids to produce xanthoxin (the first C15 intermediate); and AAO catalyses the final step of ABA biosynthesis, which converts ABA aldehyde to ABA. [126]

Half of the sentences (i.e. 50 describing interactions, 50 describing no interactions) were used to adjust the weightings of the various scoring components described above in order to find a good combination for differentiating between true positives (i.e. true interactions with entities correctly assigned) and false positives (i.e. non-interactions, or interactions with entities mis-assigned). The weightings were adjusted manually by assessing false positives and negatives and adjusting accordingly. For instance, if it was found that most false positives generally had more words in between metabolite entities than true positives, then the penalty for each word between the first and last entity would be increased and the algorithm re-run. A number of iterations of adjustments were made, but no attempt was made to highly optimize the choice of weightings and thresholds, as the sample size of sentences was relatively small and unlikely to be highly representative of relevant literature as a whole. The effectiveness of the chosen weightings was evaluated using the remaining set of 100 sentences.

13.2 Evaluation pathways

Rather than create a set of manually-annotated sentences or abstracts to evaluate the method, performance against manually-curated pathways in the EcoCyc database was assessed. This is a similar approach to that adopted by Yuryev *et al*. [40] in the context of automated signaling pathway construction and Rodríguez-Penagos *et al*. [42] when evaluating the automated reconstruction of a bacterial regulatory network.

Three pathways were chosen from EcoCyc and the original papers cited in each of these EcoCyc entries collected. The three pathways are shown in Figure 5: the pantothenate and coenzyme A biosynthesis pathway (8 papers), the tetrahydrofolate biosynthesis pathway (13 papers) and the aerobic fatty acid β-oxidation I pathway (11 papers). All three pathways are from *E. coli K-12 substr. MG1655*. All reactions in all three pathways have at least one substrate,

product and enzyme; some reactions have multiple substrates and/or products, but there is never more than one enzyme.

I chose to annotate only the Abstract and Introduction of the referenced papers using the metabolic reaction system and compare the results to the relevant pathways within EcoCyc. The decision to exclude the Methods, Results and Discussion sections was in part a pragmatic one (it reduced the amount of text that was needed to be examined manually in order to evaluate the performance of our system), but was also guided by previous research concerning the information content of the different sections of full-text articles. For example, Shah *et al*. [127] undertook an analysis of the distribution of protein and gene names in 104 articles, and concluded that the Abstract and Introduction were the best sources of information about entities and their interactions, with the Methods and, to a lesser extent, the Results sections often proving problematic (for example, keywords unique to the Methods section commonly refer to reagents and experimental techniques).

13.3 Measuring performance

To gain a rounded picture of how well the system performs, I considered the quality of its predictions for different aspects of our evaluation data: the entities (enzymes, small molecules) within a pathway; the metabolic reactions within a pathway; the binary relationships (enzyme-substrate, enzyme-product, substrate-product) within a reaction; and whole pathways.

Given that predictions were compared to manually-curated pathways, rather than to gold-standard corpus annotations, we chose to adopt a similar approach to measuring performance to that of Rodríguez-Penagos *et al*. [42]. However, in a preliminary evaluation of entity tagger performance, gold-standard manually-annotated corpora were used, rather than curated pathways. In this context I was able to calculate the standard recall, precision and F-score metrics used in the majority of text mining research. Consequently the main performance measures used are:

Recall(C) Of the reactions/relationships/entities within a corpus of texts, the percentage that have been extracted — here "C" stands for "corpus".

Recall(P) Of the reactions/relationships/entities within a manually-annotated pathway, the percentage that have been extracted — here "P" stands for "pathway".

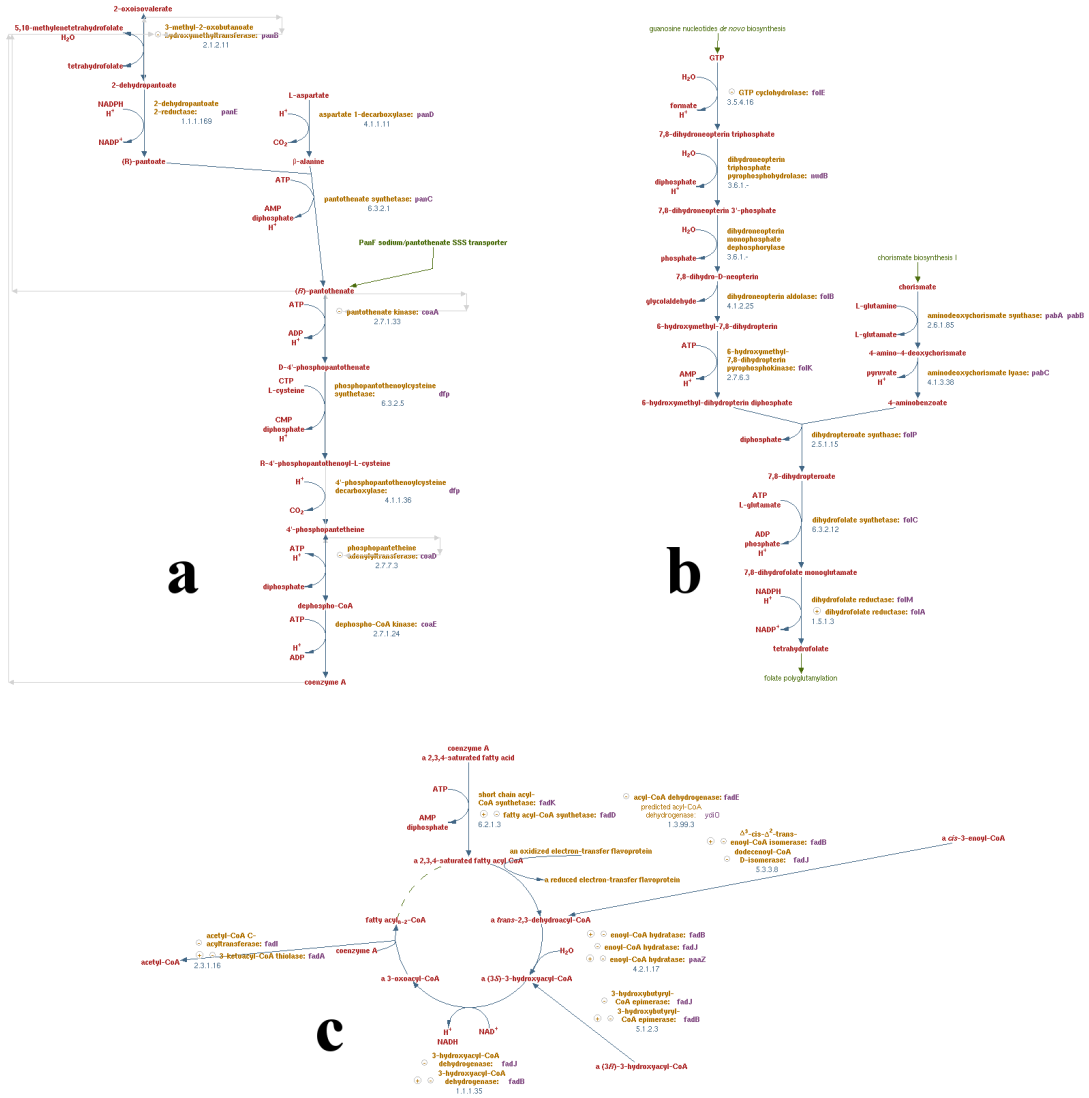

Figure 5: The three chosen pathways from EcoCyc used for evaluation of the metabolic reaction extraction algorithm: a) pantothenate and coenzyme A biosynthesis; b) tetrahydrofolate biosynthesis; and c) aerobic fatty acid β-oxidation I.

Precision Of the extracted reactions/relationships/entities, the percentage that are correct.

F-score The harmonic mean of recall(C) and precision.

Note that, in evaluating recall, only the primary metabolites belonging to the main route along the metabolic pathway were taken into account. Hence side metabolites, such as $ATP \rightarrow ADP + P_i$, are ignored. This approach was taken because it is common practice for authors to omit details about side metabolites from published papers, leaving them to be inferred by the reader.

When judging the accuracy of named entity taggers, there is a choice to be made between "strict" matching criteria (where the tagger is required to match a given name exactly) and "sloppy" matching criteria (where the tagger is not required to match the name boundaries exactly to score a "hit"; any sized overlap between the gold-standard entity and the tagged entity will count as a match). For example, consider the following tagged sentence fragment:

```
...is a key precursor of the <molecule>4′-phosphopantetheine</molecule>
moiety of...
```

Using sloppy matching criteria, credit is given for annotating *phosphopantetheine*, *4′-phosphopantetheine* or *4′-phosphopantetheine moiety*, but also for *key precursor of the 4′*; whereas strict matching criteria require an exact match to *4′-phosphopantetheine*.

In this research I adopted sloppy matching criteria on the grounds that they have proved more informative than strict criteria in the context of gene/protein NER in general, and of gene/protein relationship extraction in particular. With respect to NER, in the vast majority of cases where a match was found using sloppy criteria but not with strict criteria, the core part of the entity name was correctly identified [128]. Strict criteria were deemed misleading because they are highly sensitive to the essentially arbitrary choices made when drawing up annotation guidelines for the evaluation corpora — for example, whether the word *mouse* is part of the protein name in the phrase *mouse oxytocin*. With respect to NER in the specific context of relationship extraction, a manually corrected F-score was only 4 percentage points lower than the sloppy F-score, but 20 points greater than the strict F-score [128]. The manually corrected F-score was calculated by the manual assessment of each tagged entity that was counted as a miss using strict criteria, but a hit by sloppy criteria.

In the data sets used for this research there are a few examples where sloppy matching

criteria arguably give a misleading impression about how well a complex entity name has been tagged. With sloppy matching, both the following examples of sub-optimal tagging score a "hit":

- The significant truncation of the long entity name *Geranyl pyrophosphate:(-)-endo-fenchol cyclase* to *endo-fenchol cyclase*;
- The splitting of the single entity *TPS-d3 family members of conifer diterpene synthases* into the two tagged entities *TPS-d3* and *conifer diterpene synthases*.

However, such examples were comparatively rare, and it was concluded that the number of false negatives that appear to be true negatives with strict criteria is a more significant problem than the number of false positives that appear to be true positives with sloppy criteria.

14 Results

14.1 Pre-evaluation of entity taggers

I performed a preliminary evaluation of the performance of BANNER and OSCAR3 on the GENIA corpus [84], which contains 2,000 biomedical abstracts related to the specific topic of human blood cell transcription factors. GENIA was chosen because it contains annotations for a broad range of biological and chemical entities. Additionally OSCAR3 was tested using the dedicated Fraunhofer SCAI chemical corpus [91], which contains 101 abstracts from chemistry papers. Neither tool was developed using either of these corpora: BANNER was trained on the BioCreative corpus [129], and OSCAR3 was trained on two corpora of full-text articles from RSC journals and abstracts [94].

BANNER scored 72% for precision, recall(C) and F-score on GENIA. This is roughly in line with expectations; it has been previously shown that a range of protein/gene name taggers perform less well on GENIA than on some other widely-used corpora, and that this is (at least in part) attributable to the chosen annotation criteria (see, for example, the analysis in [130]).

The results for OSCAR3 are more interesting and are presented in Figure 6. Two features stand out from these results: the best performance of OSCAR3 on both corpora is worse than had been expected from results presented elsewhere [56], with peak F-scores of 62% and 48%

on the Fraunhofer SCAI corpus (Figure 6a) and the GENIA corpus (Figure 6b) respectively; and the performance on the GENIA corpus is significantly worse than that on Fraunhofer SCAI.

A preliminary examination of the tagged text generated by OSCAR3 for both corpora indicated that a significant proportion of the false positives were attributable to acronyms being tagged as the names of chemicals. This is a known problem (identified in the original OSCAR3 paper by Corbett & Murray-Rust [93]) and one that the authors advocate addressing at the level of the wider text-mining framework.

In this spirit, a simple method for resolving acronyms was developed. Any putative acronym (i.e. any uppercase token of more than one letter) is deemed to be a false positive unless either a) a defining chemical name is found in the text preceding it, or b) OSCAR3 gives it a confidence score of 0.5 or more. The latter criterion is used to allow for the presence of commonly occurring molecules for which acronyms are frequently used without explicit definition (e.g. *NAD*). This approach achieved a significant improvement in precision at the cost of a negligible drop in recall (Figure 6c). Bearing these results in mind, we henceforth used OSCAR3 with the threshold set to zero, thereby maximizing recall.

Our training corpus of sentences containing the names of at least two small molecules (see Section 60) was also used to assess whether, in cases where BANNER tags multiple protein names within a single sentence, it is advantageous to prefer names that end in *-ase* or *-ases*. Of the 77 enzyme names in the training corpus, 60 end in *-ase(s)*. As expected, the suffix *-ases* commonly occurs when a text refers to a class of enzymes in general, whereas the suffix *-ase* is used when a specific enzyme is being discussed in the context of a particular reaction. I also found, however, that the naming of multiple proteins in a sentence describing a metabolic reaction was uncommon. Therefore, this approach was not incorporated into the current method.

14.2 Performance of entity taggers on metabolic corpora

I began by undertaking a standard analysis of tagger performance by evaluating BANNER and OSCAR3's scores for all the entities in the Abstract and Introduction of each of the papers associated with the three evaluation pathways. All protein and small molecule names were manually annotated to achieve a true recall figure. Results are shown in Table 6.

Performance here is significantly higher than it was for GENIA. It is worth noting that, in the case of BANNER, the performance on this corpus is very similar to its performance on

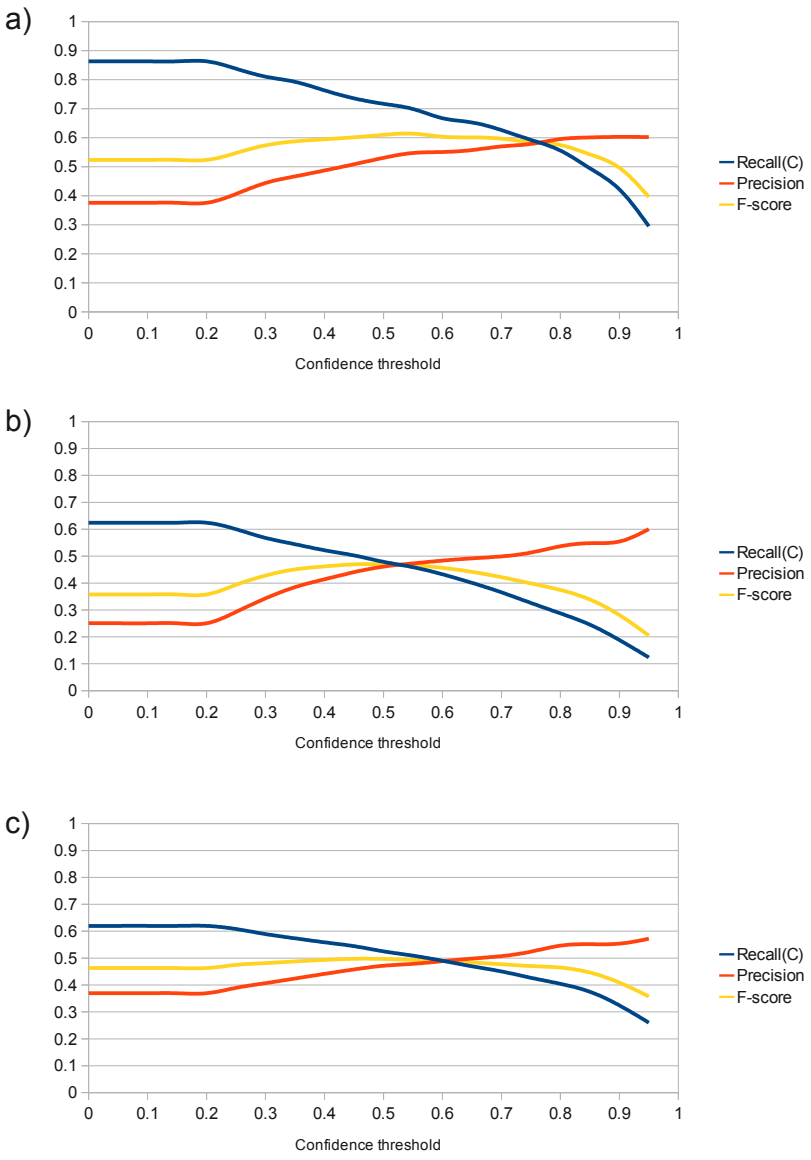

Figure 6: Graphs showing the performance of OSCAR3 at a range of confidence thresholds. Performance is shown under the following conditions: a) when applied to the SCAI chemical corpus; b) when applied to the GENIA corpus without acronym detection; and c) when applied to the GENIA corpus with acronym detection. The y-axis gives the recall(C), precision and F-score values in the range 0 to 1.

	Protein names tagged by BANNER	Small molecule names tagged by OSCAR3
Pantothenate and coenzyme A biosynthesis pathway		
Recall(C) (%)	81 (112/139)	96 (329/343)
Precision (%)	85 (112/132)	86 (329/384)
F-score (%)	83	91
Tetrahydrofolate biosynthesis pathway		
Recall(C) (%)	93 (250/268)	82 (528/647)
Precision (%)	76 (250/327)	95 (528/558)
F-score (%)	84	88
Aerobic fatty acid β-oxidation I pathway		
Recall(C) (%)	91 (341/376)	81 (456/565)
Precision (%)	82 (341/414)	92 (456/494)
F-score (%)	86	86

Table 6: The tagging performance of BANNER and OSCAR3. The tagging performance of the NER tools when applied to the Abstracts and Introductions from papers referenced in EcoCyc with respect to the three evaluation pathways. Taking the BANNER column for the pantothenate and coenzyme A biosynthesis pathway as an example, the numbers in brackets indicate that BANNER correctly identified 112 out of the 139 protein names (recall row); and of the 132 names it tagged, 112 were correct (precision row). The OSCAR3 results are with a confidence threshold of zero.

gene/protein interaction corpora such as AIMed [131] and the LLL training corpus created for the 2005 LLL challenge [132], for which F-scores of 82.9% and 84.1% were reported in [1].

14.3 Relationship extraction

The reactions extracted from the three EcoCyc pathways were evaluated in a number of different ways. In the first evaluation, reactions were considered correct if all primary substrates, products and the enzyme were extracted correctly. In the second, the ability to extract the enzyme was not taken into account. The results of these two evaluations can be seen in Table 7. The precision scores show the percentage of extractions that were correct under the previously described criteria. The recall(P) scores show the percentage of the reactions in the pathway that were successfully extracted. As was described in Section 13.3, recall(P) is not the true recall (which would be the percentage of metabolic reactions in the source documents that were successfully extracted) and as such the F-score cannot be calculated.

A third evaluation (see Table 8) broke the reactions down into binary interactions (substrate-product, substrate-enzyme and product-enzyme) that allowed a more granular evaluation. Consider the following reaction catalysed by the enzyme *E1*:

$$S1 + S2 \rightarrow P1$$

The reaction can be broken down into the following binary interactions:

- S1 — P1
- S2 — P1
- E1 — S1
- E1 — S2
- E1 — P1

If one of the substrates were missed in the extraction, the extraction would be considered incorrect using the first evaluation criteria considering whole reactions, but the second criteria

	Correct reactions (ignoring enzyme)	Correct (including enzyme)
Pantothenate and coenzyme A biosynthesis pathway		
Recall(P) (%)	78 (7/9)	56 (5/9)
Precision (%)	59 (24/41)	41 (17/41)
Tetrahydrofolate biosynthesis pathway		
Recall(P) (%)	90 (9/10)	70 (7/10)
Precision (%)	60 (39/65)	38 (25/65)
Aerobic fatty acid β-oxidation I pathway		
Recall(P) (%)	29 (2/7)	29 (2/7)
Precision (%)	30 (11/37)	14 (5/37)

Table 7: The performance of the metabolic reaction extraction method on the three evaluation pathways. Taking the "correct reactions (ignoring enzymes)" column for the "pantothenate and coenzyme A biosynthesis" pathway as an example, the numbers in brackets indicate that the algorithm correctly identified 7 out of the 9 reactions in the curated EcoCyc pathway (recall row), giving 78%; and of the 41 identified interactions (precision row), 24 were valid reactions (irrespective of whether they belong to the pathway or not), giving 59%. A reaction for which the substrate(s) and product(s) have been correctly assigned, but not the enzyme, is deemed correct in column two, but incorrect in column three.

evaluating binary interactions would recognise the extraction as partly correct. This evaluation of binary interactions also allows the results to be compared to tools used for the extraction of binary protein-protein interactions (see Table 9).

A visual summary of the complete set of results for the three pathways is given in Figures 7-9.

Fair and meaningful comparisons within the field of biological text mining are extremely difficult; for example, a single system may give a wide range of different performances even when applied to different corpora within the same sub-domain. In this research, a prominent feature of the results (as presented in Tables 7 and 8) is that the algorithm performs noticeably less well on the aerobic "fatty acid β-oxidation I" pathway than on the other two pathways. To a significant extent this appears to be attributable to the distinctive ways that reactions in fatty acid pathways are commonly described, for example in terms of molecular addition (with no explicit product mentioned):

> Enoyl-CoA hydratase catalyzes the second reaction of the fatty acid β-oxidation, i.e., the syn addition of water to α, β-unsaturated fatty acyl-CoA thioesters.

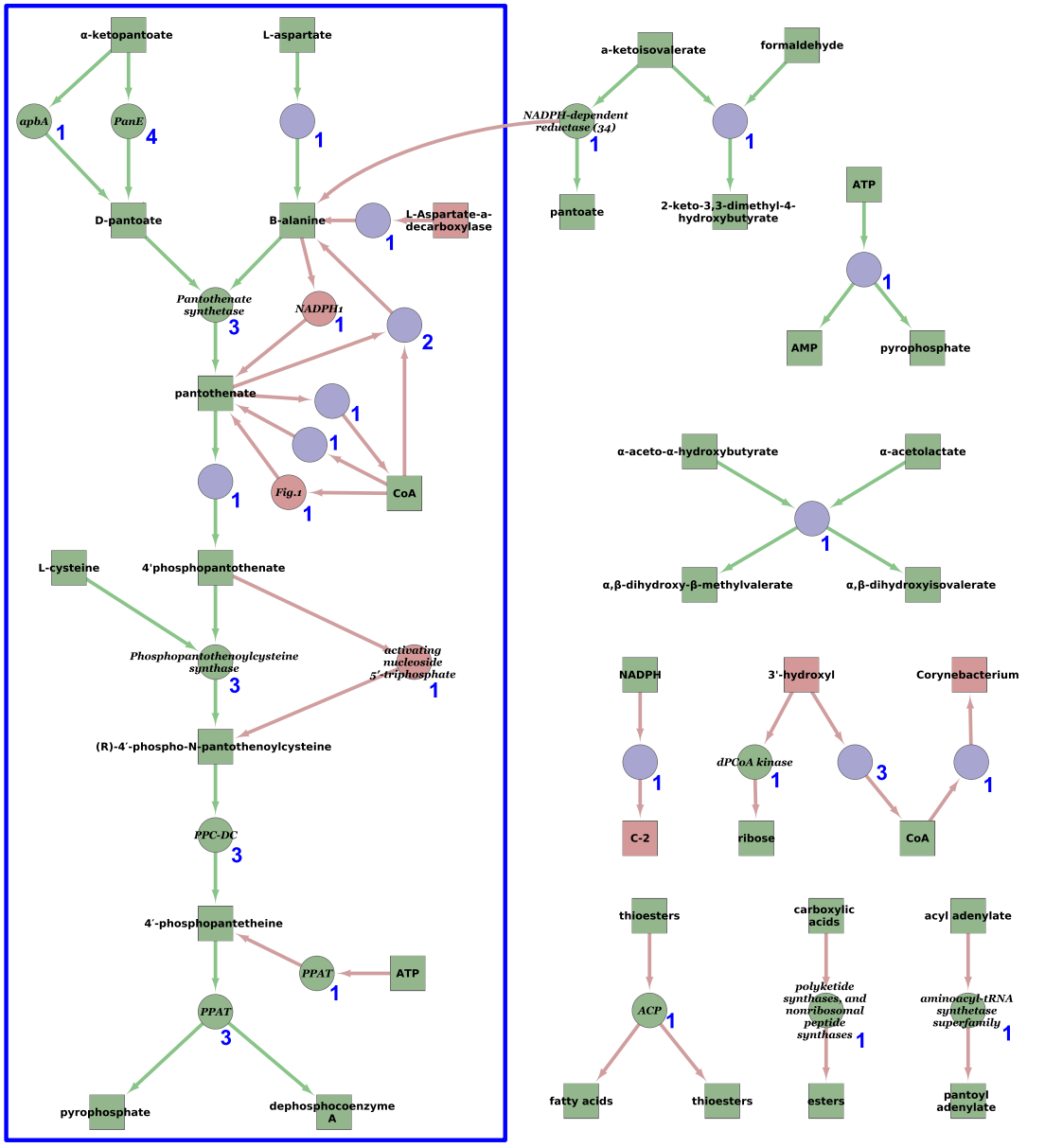

Figure 7: The reconstructed pantothenate and coenzyme A biosynthesis pathway from mined reactions. Squares are small molecules, circles are enzymes, and a pair of arrows is used to denote a single reaction (the first for the interaction substrate-enzyme, and the second for the interaction enzyme-product). Items labeled green are correct, items labeled red are incorrect, and a purple circle denotes an extracted reaction with no corresponding enzyme extraction. The number next to a reaction indicates the number of times that reaction was extracted from the set of source texts. The reactions on the right-hand side of the figure (lying outside the blue rectangle) are reactions extracted by our algorithm that are not part of the manually-annotated pantothenate and coenzyme A biosynthesis pathway from EcoCyc given in Figure 5a.

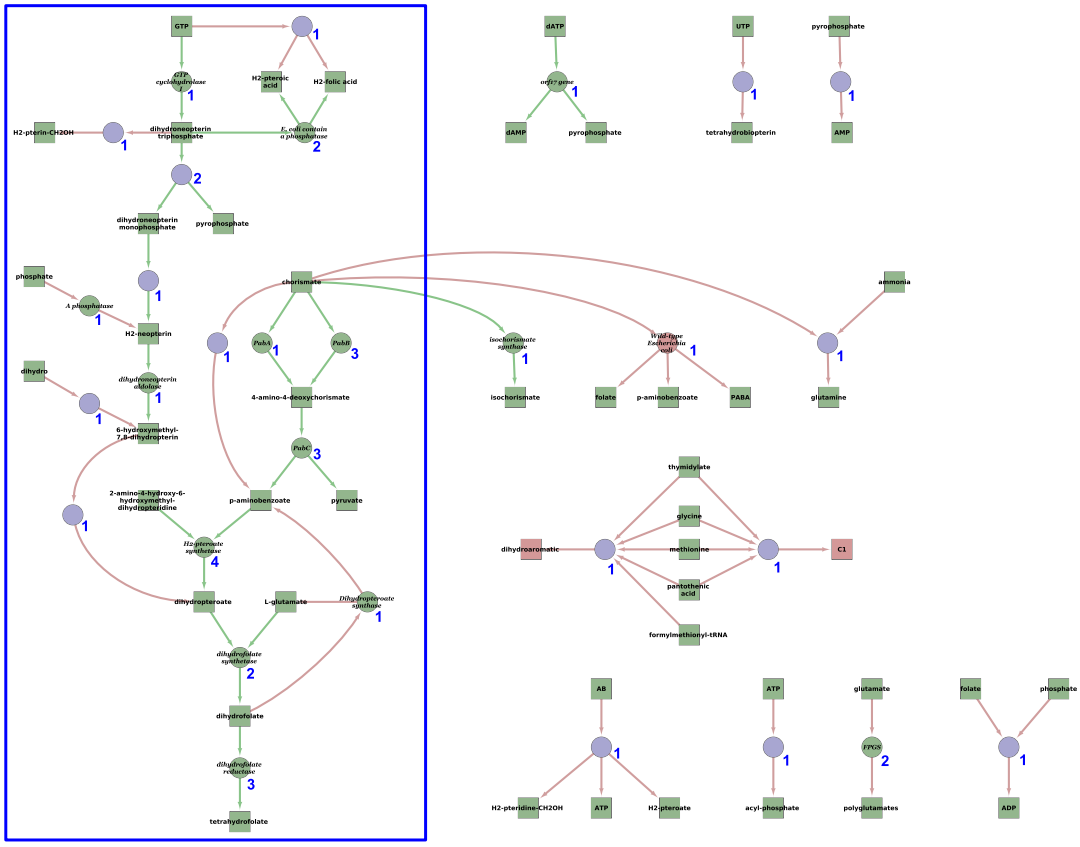

Figure 8: The reconstructed tetrahydrofolate biosynthesis pathway from mined reactions. The network is structured in the same way as Figure 7 on page 71. The reactions on the right-hand side of the figure (lying outside the blue rectangle) are reactions extracted by our algorithm that are not part of the manually-annotated pantothenate and coenzyme A biosynthesis pathway from EcoCyc given in Figure 5b.

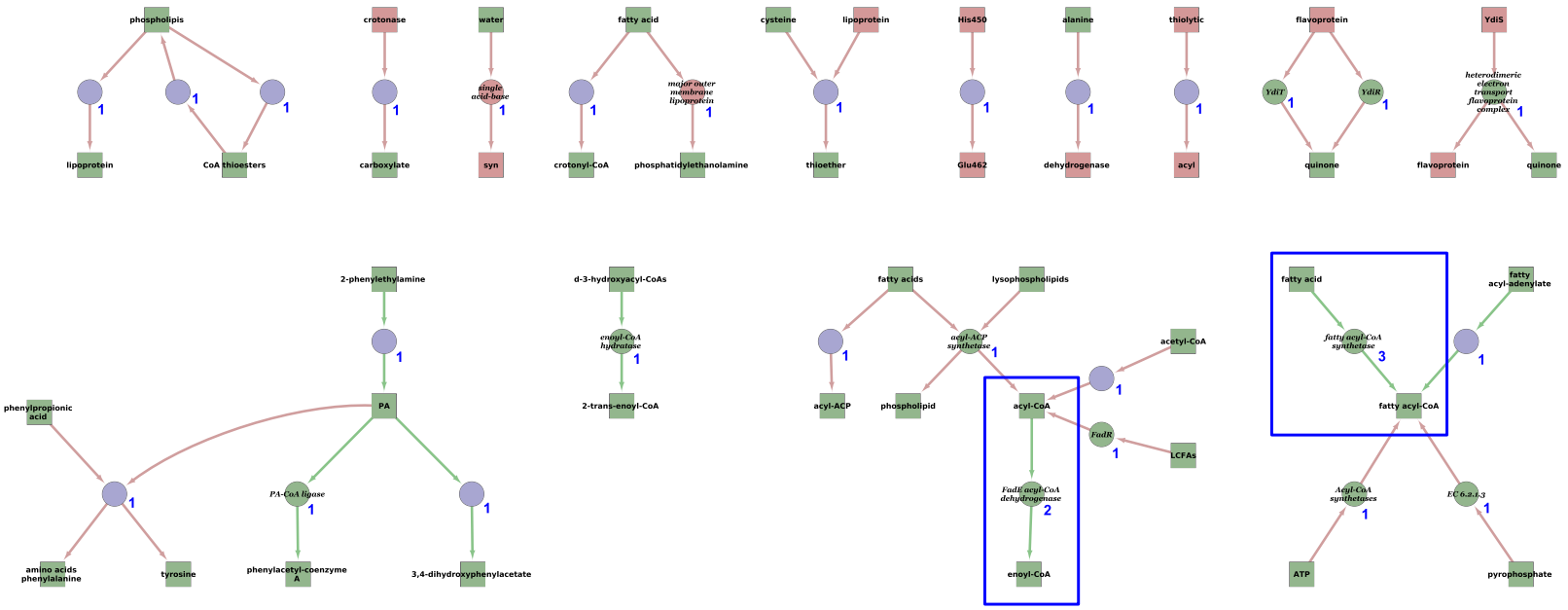

Figure 9: The reconstructed aerobic fatty acid β-oxidation I pathway from mined reactions. The network is structured in the same way as Figure 7 on page 71. The reactions on the right-hand side of the figure (lying outside the blue rectangle) are reactions extracted by our algorithm that are not part of the manually-annotated pantothenate and coenzyme A biosynthesis pathway from EcoCyc given in Figure 5c.

	Substrate-product	Substrate-enzyme	Product-enzyme	Total
Pantothenate and coenzyme A biosynthesis pathway				
Recall(P) (%)	67 (10/15)	58 (7/12)	55 (6/11)	61 (23/38)
Precision (%)	59 (35/59)	65 (13/20)	59 (13/22)	60 (61/101)
Tetrahydrofolate biosynthesis pathway				
Recall(P) (%)	82 (9/11)	64 (7/11)	70 (7/10)	78 (25/32)
Precision (%)	48 (55/114)	62 (28/45)	58 (26/45)	53 (109/204)
Aerobic fatty acid β-oxidation I pathway				
Recall(P) (%)	20 (2/10)	38 (3/8)	38 (3/8)	31 (8/26)
Precision (%)	40 (12/30)	80 (8/10)	67 (6/9)	53 (26/49)

Table 8: Binary interaction extraction performance for all three evaluation pathways. Numbers in brackets were calculated as for Table 7.

However, in the absence of a substantially larger data set, it is not possible to draw firm conclusions.

Notwithstanding these caveats and challenges, note (with considerable caution) that these results appear to be somewhat better than those achieved using the EMPathIE system [46]. However, no direct comparison is possible.

Caution should, of course, be exercised when making comparisons between different sub-domains and where the evaluation strategies are different — particularly as our evaluation did not involve a calculation of the true recall, but rather the recall of the relevant pathways. Nevertheless, it is useful to consider how the performance of this method for extracting metabolic reactions compares to that in the well-studied sub-domain of gene/protein interaction extraction. Here (in Table 9) the performance of this method is briefly compared with the reported performance of three contrasting gene/protein interaction tools: the rule-based RelEx method [133], which was the best-performing method in the evaluation reported in [1]; the NLP tool AkanePPI [134] trained on the BioInfer corpus [135]; and the simple baseline(k) algorithm in [1].

The analysis of extracted reactions as ternary relationships (in Table 7) and as binary relationships (in Table 8) suggest both that the method performs reasonably well when placed in the wider context of biomedical relationship extraction, and that metabolic reaction extraction is more tractable than has hitherto been assumed.

Method	Interaction type	Range of scores on different corpora (%)	
		Precision	Recall
RelEx	Protein-protein	39-80	45-72
Baseline(k)	Protein-protein	23-54	52-67
AkanePPI (trained on BioInfer)	Protein-protein	29-77	40-56
Method described in this paper	Substrate-product	40-59	20-82
Method described in this paper	Substrate-enzyme	62-80	38-64
Method described in this paper	Product-enzyme	58-67	38-70

Table 9: Comparison of the performance of methods for extracting gene/protein interactions with that of the method for extracting metabolic reactions presented here. The range of scores for the gene/protein extraction tools are for five corpora as evaluated in [1]. The scores for this metabolic reaction extraction method summarize those in table 8, i.e. they are broken down into the same three binary interactions and the range is for the three evaluation corpora.

15 Discussion

Here I have presented a simple method for extracting metabolic reactions from free text. I have shown that it successfully extracted a high percentage of reactions for two out of three pathways; the third pathway, dealing with fatty acid metabolism, proved particularly challenging owing to the distinctive way in which reactions are described (for example, in terms of molecular addition). Insofar as comparisons with broadly comparable methods are possible, it appears that this approach performs rather well; that, at least, is what the brief comparison with the performance of gene/protein interaction extraction methods suggests, with both precision and recall at comparable levels.

Given that information about secondary metabolites such as ATP is frequently omitted from source papers, I have focused on the extraction of primary metabolites, rather than side metabolites, in the evaluations presented here. Clearly, this lack of information about side metabolites in the literature is an obstacle to the fully automated construction of complete metabolic pathways using text-mining methods. However, a more realistic goal for a metabolic text mining system is to support manual curation. In this latter context, I believe these evaluations show that this method could prove immediately useful to database curators, who are already used to having to infer the side metabolites when metabolic reactions are incompletely specified in the literature. It is important to remember, however, that the evaluation was car-

ried out by extracting reactions from articles known to be relevant to the pathway — resulting in relatively few irrelevant reactions being extracted. In the next part, LiMPET is tasked with extracting reactions from larger sets of articles with varying levels of relevance which results in relevant reactions often being far out-numbered by irrelevant reactions. Methods for calculating the relevance of extractions are also discussed.

There are a number of ways that the method could be improved, for example by incorporating techniques for handling negation (and speculation) and resolving anaphora, and the system might benefit from using more sophisticated tool in place of our present simple acronym resolution strategy, such as the widely-used Acronym Resolving General Heuristic (ARGH) program [136].

But perhaps more interesting is the fact that this relatively simple method performs so well, especially in light of prior assumptions that this is a particularly challenging sub-domain. There are several reasons why this may be the case:

- Whole reactions are commonly described in a single sentence.
- A single sentence commonly describes a single reaction and nothing else.
- Entity taggers appear to be reasonably accurate in a metabolic context, with most enzyme names having the suffix *-ase* or *-ases*.
- Keyword lists appear reasonably discriminatory when distinguishing metabolites from non-metabolites and substrates from products.
- Most reactions are described multiple times in the literature; typically at least one occurrence will be worded in such a way that the information is relatively easy to extract.

Part IV

LiMPET — a metabolic pathway extraction pipeline

16 Introduction

In the previous part I described the development of a metabolic reaction extraction algorithm. The algorithm forms the local association analysis component of LiMPET — attempting to determine reactions given the contents and structure of individual sentences. In the following part I will describe the global association analysis components of LiMPET — the integration of extractions from different sources and publicly available data.

The bulk of this part is concerned with constructing a pathway from the individual extractions made by the core algorithm and determining the correct and relevant parts of the returned network. The first challenge in constructing a pathway from individual extraction is merging together different mentions of the same reaction. As has been discussed previously, it is sometimes difficult to define what constitutes a metabolic reaction description. Some descriptions may just include the primary metabolites while another includes the enzyme and another still includes side reactions. This makes the consolidation of metabolic pathway data from different sources inherently difficult as it can be challenging to determine the 'important' parts of the reaction on which to base any merges. Moreover, as a text mining project there will always be a certain level of unreliability in the data being consolidated and it is important not to let this affect the integrity of correct extractions.

The second challenge in pathway building is correctly linking reactions together such that the product of one reaction is the substrate of the next. Firstly, different names for the same metabolite must be disambiguated due the many different names that a single metabolite must have. Secondly, a determination must be made on the appropriate metabolites through which reactions can be linked. Linking together all reactions containing acetyl-CoA is unlikely to yield a useful pathway.

In approaching both of these challenges one must also consider the host organism of the ex-

tracted reactions, especially when a wide range of literature is being mined — not just literature that is known to be relevant. In the spirit of the core algorithm of LiMPET I have attempted to solve these challenges using heuristic algorithms and already available tools.

Also in this part I describe the development of a component to allow the automated retrieval of relevant articles to mine (see Section 17.1). While in the previous part the core algorithm was tested on collections containing less than 20 articles, this component allows the mining of hundreds of articles. This introduces its own problems, however, resulting in the extraction of far more incorrect and irrelevant reactions. In Section 17.6.2 I describe a method for scoring the correctness and the relevance of extracted reactions to allow the appropriate data to be found.

17 Pipeline components

As input the pipeline takes a list of MetaCyc pathway IDs and the NCBI Taxonomy ID of the organism of interest. For instance, consider a user interested in alanine biosynthesis in *Mycobacterium tuberculosis*. The user would search MetaCyc using the term "alanine biosynthesis" and find three alternative pathways with the designations I, II and III. The ID can be retrieved from the URL for each of the alternative pathways:

```
http://www.metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=ALANINE-VALINESYN-PWY
```

The IDs are supplied to LiMPET as a vertical bar delimited list:

```
ALANINE-VALINESYN-PWY|ALANINE-SYN2-PWY|PWY0-1021
```

These pathways are merged and are used as a seed pathway for the retrieval of relevant articles. While it would be more user-friendly for the user to simply supply the term "alanine biosynthesis", it is impossible to search MetaCyc using pathway names using the API.

The organism ID can be obtained by searching NCBI Taxonomy with the ID being clearly displayed on an organism's page. In the case of *M. tuberculosis*, the ID "1773" would be supplied to LiMPET.

17.1 Literature search

A set of literature search queries is generated from the inputted MetaCyc pathways and organism. One search query is constructed for each unique metabolite in the seed MetaCyc pathway,

containing synonyms for the metabolite (obtained from the ChEBI database) and the organism of interest (obtained from the NCBI Taxonomy database). The following query would be constructed for the metabolite *pyruvate* and the organism *M. tuberculosis* (truncated lists of synonyms are used for presentation purposes):

```
(("M.tuberculosis"[All Fields]) OR ("Mycobacterium tuberculosis"[All
Fields]) OR ("Bacterium tuberculosis"[All Fields])) AND (("pyruvate"[All
Fields]) OR ("pyruvic acid"[All Fields]) OR ("alpha-ketopropionic
acid"[All Fields]))
```

The boolean terms `OR` and `AND` ensure that the retrieved document records contain any organism synonym and any metabolite synonym.

Organism synonyms are retrieved using the dictionary provided by the LINNAEUS organism named entity recognition library (which is based on the NCBI Taxonomy database) [111]. Metabolite synonyms are retrieved from a local ChEBI database [103]. Not all metabolites are included in the query — currency molecules, such as *ATP* and *ADP*, are excluded as they are not meaningful in the identification of pathways and may lead to the retrieval of many irrelevant articles.

LiMPET uses PubMed to carry out literature searches because it is the *de facto* standard for life science research and has a stable, mature API (named NCBI E-Utils) [2]. It is important to be aware of its limitations, however. By default, PubMed returns results in reverse chronological order. Ordering by relevance was evaluated, but no performance improvement was found. By default, LiMPET retrieves the 100 most recent articles. This was determined to provide a good balance between processing time and literature coverage, although, with particularly well-studied organisms there may be far more articles than can feasibly be mined using the algorithm. While there were difficulties parallelising the tool, this approach would significantly reduce the processing time as each article in a retrieved set would be mined independently.

PubMed does not have access to the full-text of papers and is limited to searching the title and abstract. Google Scholar, on the other hand, indexes the full-text of articles, but, unlike PubMed, provides no API. While third-party libraries have been developed to allow automated searching using Google Scholar, such libraries retrieve results by parsing the specific website HTML. A relatively minor change to the markup, therefore, could result in broken functionality. This limits their long-term stability and the implementation of such libraries in a

public tool would, therefore, be inappropriate. Using PubMed, it must be taken into consideration that information that is buried in the full-text, with no indication in the abstract, will be missed.

17.2 Literature retrieval

Text-mining tools have traditionally focused on the mining of abstracts as they can be retrieved using a stable API (or as a bulk download) in a consistent format. As the PMC Open-Access Subset has grown in recent years, text-mining research has begun to utilise the subset's full-text articles in the evaluation of new methods. As discussed in Section 5, however, the subset only contains a small percentage of all published research. While the subset is useful for testing certain text-mining methods it is not particularly useful for the end user[8].

This lack of automated access to full-text literature was first approached as a technical challenge (although, as described in Section 5, the problem was revealed to be one of politics and legalities). In addition to metadata regarding an article, records obtained by the PubMed API include links to the location of the article on publishers' websites. A system was developed which followed these links and employed a general screen-scraping method (independent of any specific markup) to retrieve the full-text article. This was a complex process as the links supplied by PubMed do not have a standard destination. They may link to the abstract or directly to the full-text. The publisher may hold the full-text in HTML and/or PDF, and the PDF may be linked to directly or be displayed in a frame within a webpage.

The evaluation of LiMPET in this chapter was carried out using full-text articles retrieved using this method. As will be discussed in Part VI, however, this behaviour cannot be made available in the public tool. Therefore, the public release of LiMPET only has the ability to retrieve abstracts and full-text articles from the PMC Open-Access Subset using the NCBI E-Utils API (the implications of which are shown in Section 18.2).

17.3 Metabolic reaction extraction

Metabolic reactions were extracted using the algorithm described in Part III. While minor bugs were addressed throughout LiMPET's development, the core methodology remained static.

[8]The usefulness of abstracts and open-access PMC articles in this domain is investigated in Section 18.2.

17.4 Assignment to organisms

While the literature search strategy should retrieve articles relevant to a specific organism, it remains necessary to assign individual reactions to an organism as articles rarely mention just a single organism. The organism of interest may be mentioned in passing in an article abstract while the article deals principally with a different organism. Reactions in such an article should not be assigned to the organism of interest. An article dealing with the organism of interest may compare against reactions in other organisms. Such reactions should be recognised as not belonging to the organism of interest.

A simple, heuristic approach to this problem was developed — similar to entries in the gene normalisation task of BioCreative III [45]. Using a development corpus of 30 documents the following rules were developed, in order of priority:

1. If an organism is mentioned within a reaction sentence, the reaction is associated with this organism. If multiple organisms are mentioned, the reaction is assigned to all.

2. If the previous point does not apply, the reaction described will belong to the first organism mentioned in the paper.

This small number of rules and their simplicity was surprising considering the expectation that this may be a very challenging problem. Nevertheless, an algorithm incorporating these rules was developed and tested on a corpus of 20 papers. Of the 78 reactions described in the test set, 63 were assigned to the correct organism for an accuracy of 81%. By simply assigning each reaction to the first organism mentioned in the paper, only 51 were assigned to the correct organism for an accuracy of 65%.

17.5 Pathway building

Up to this point I have described the extraction of single metabolic reactions. There will be cases, however, where the same reaction is extracted from multiple source sentences. If left as separate reactions, the outputted pathway would likely be very hard to understand. Merging separate extractions also provides an avenue to score the correctness of putative reactions (see Section 17.6.1).

Metabolic reactions do not exist in isolation, but instead form pathways of reactions that have evolved to carry out a specific function, such as the breakdown of large molecule in mul-

tiple steps in order to harvest energy for the cell. In this section I discuss the methods I have employed to link together individual metabolic reactions in the face of many difficulties, such as promiscuous molecules that are involved in many reactions and incorrectly extracted reactions.

To merge and link reactions together to form pathways it is necessary to disambiguate small molecule names. LiMPET does so by obtaining InChIs from ChEBI [103] (Section 9.3.2 contains a discussion of small molecule identifiers and chemical databases). As an InChI needs to be retrieved for each small molecule extracted from the text, using the ChEBI web services would be unwise due to usage limits and the time required for a large number of queries. A local copy was, therefore, downloaded and the data stored in an embedded SQL database within LiMPET. Variants of all synonyms (such as disregarding stereochemistry and whitespace; see Appendix II for a full explanation) were pregenerated and indexed to allow for quick look-up. In the program output, entity mentions assigned the same InChI identifier are merged to create a single entity regardless of the individual extracted names.

Merging reaction extractions is not straightforward, however. Two extractions that contain the exact same substrates and products can be safely merged, but in cases where some, but not all, metabolites are shared by two extractions, a decision must be made about whether and how to merge. For instance, consider the reactions:

$$A \rightarrow C$$

$$A + B \rightarrow C$$

There are two different ways of merging the reactions: taking the union (creating a reaction including all metabolites from all the merged reactions) or taking the intersection (creating a reaction including only those metabolites found in all the merged reactions). If we assume that both reactions are correctly extracted, it is typically safe to assume that B is a side metabolite and is not always referenced when describing the reaction — in which case, the union of the reactions should be taken. Extractions are not always correct, however, so it may be that B is an erroneous addition and is not involved in the reaction — in which case, the intersection should be taken. Initially an algorithm was developed that would take the union of two extracted

reactions (as the intersection would result in the potential loss of information) if the reactions had at least one substrate and at least one product in common. Therefore, the reactions above would be merged to create a single reaction ($A + B \rightarrow C$) as both reactions have A as a substrate and C as a product. This algorithm was problematic, however, and caused over-merging. Consider the following three reactions:

$$A \rightarrow C$$
$$B \rightarrow D$$
$$A + B \rightarrow C + D$$

The algorithm would result in the union of all three reactions to create $A + B \rightarrow C + D$. If the third reaction extraction was incorrect, however, this would effectively hide $A \rightarrow C$ and $B \rightarrow D$ as separate reactions from the user. In this case, the three reactions should simply remain unmerged.

Ideally, a statistical method would be used that accounted for the number of times a reaction was extracted. If $A \rightarrow C$ was extracted 10 times and $B \rightarrow D$ was extracted 10 times, while $A + B \rightarrow C + D$ was extracted only once, the reactions could be confidently left separate. Unfortunately, reactions are rarely extracted enough times for such a method to be viable.

As this over-merging proved to have a significant detrimental effect to the accuracy of outputted networks, a more conservative algorithm was employed that would only merge two extractions containing the exact same metabolites — with the exception of *currency molecules*, such as *ATP*, *ADP*, *NAD* and *NAD+*. Here *currency molecules* are defined as molecules that are not confined to a specific process and do not typically form meaningful pathways. A manually curated list of currency molecules was used (see Appendix III). Consider the following reaction extractions:

$$allantoate + H_2O \rightarrow S - ureidoglycolate + urea \quad (1)$$
$$allantoate \rightarrow S - ureidoglycolate + urea \quad (2)$$
$$allantoate \rightarrow S - ureidoglycolate \quad (3)$$

Extractions 1 and 2 would be merged as the only difference is the substrate H_2O, a currency molecule. Extraction 3 would not be merged due to the absence of the product *urea*, a non-currency molecule. This under-merging can prove problematic in cases where there are few extractions. If a specific reaction is extracted twice, but they cannot be merged, both extractions may get lost amongst the false positives. If a specific reaction is extracted four times, however, one extraction not being merged with the others has less of an impact. In practice this under-merging rarely had any significant effect on extracted pathways.

While joining reactions together to form pathways is usually trivial (i.e. if the product of one reaction has the same InChI as a substrate of a different reaction, the reactions can be joined), currency molecules can be problematic. Currency molecules tend to form a small number of highly connected nodes which the rest of the network clusters around (due to their involvement in many unrelated reactions). LiMPET, using the manually curated list of currency molecules, recognises each mention of a particular currency molecule as a unique entity. Therefore, completely separate reactions that both happen to convert *ATP* to *ADP* will not be linked together. There are problems with using a static list to identify currency molecules, however, as a metabolite's status as currency or non-currency can depend on context. While *acetyl-coenzyme A* is often confined to side reactions, it is an integral metabolite in the TCA cycle — pathways downloaded from BioCyc using the API make no distinction between "side" and "integral" metabolites.

Despite the literature search strategy (described in Section 17.1), the networks extracted are typically very large, containing false positive extractions and correctly extracted, but irrelevant, reactions in addition to the reactions relevant to the seed pathway (see Figure 12). A heuristic approach was taken to score individual metabolites for both correctness and relevance.

Metabolites were scored individually because an extracted reaction may contain both correct and incorrect information (and relevant and irrelevant information). Consider the following pathway description:

> In parasitic mode they convert the PEP generated by glycolysis to OAA, which is then reduced to malate via a cytosolic malate dehydrogenase (Figure 6). [137]

Two separate reactions are described in this sentence, but the core algorithm, unable to differentiate them, extracts the following reaction:

$PEP + OAA \rightarrow malate$

In this extraction both *OAA* and *malate* are correctly assigned to the reaction, while PEP is incorrectly assigned.

In order to score different parts of the reaction separately, LiMPET splits the reaction into binary interactions (containing a single substrate and product) — in this case $PEP \rightarrow malate$ and $OAA \rightarrow malate$. Each binary interaction is scored for correctness and relevance (see Sections 17.6.1 and 17.6.2) with the individual metabolites inheriting the highest correctness and relevance scores from a containing binary interaction. When drawing the final network a threshold is applied which would hide PEP due to its low score.

17.6 Training LiMPET

It was decided to train and test LiMPET on a particular use case: the extraction of novel pathway routes. Groups of pathways were identified in MetaCyc that showed different routes between two metabolites. For instance, consider the pathways "allantoin degradation to glyoxylate" I (from *Saccharomyces cerevisiae*) and II (from *Arabidopsis thaliana*) (see Figure 10). Both pathways begin with *allantoin* and end with *glyoxylate*, but pathway II contains an extra intermediate metabolite (*S-ureidoglycine*). LiMPET can then be tasked with extracting pathway II from the literature using pathway I as the seed pathway.

The correctness and relevance parameters were tuned using a preliminary analysis of three MetaCyc pathway extractions:

- "Allantoin degration to glyoxylate II" in *Arabidopsis thaliana* using pathway I as the seed pathway.
- "Lactose degradation II" in *Agrobacterium tumefaciens* using pathway III as the seed pathway.
- "Methylglyoxal degradation V" in *Saccharomyces cerevisiae* using pathway VII as the seed pathway

These pathways were deemed to be fairly representative of MetaCyc pathways in general in

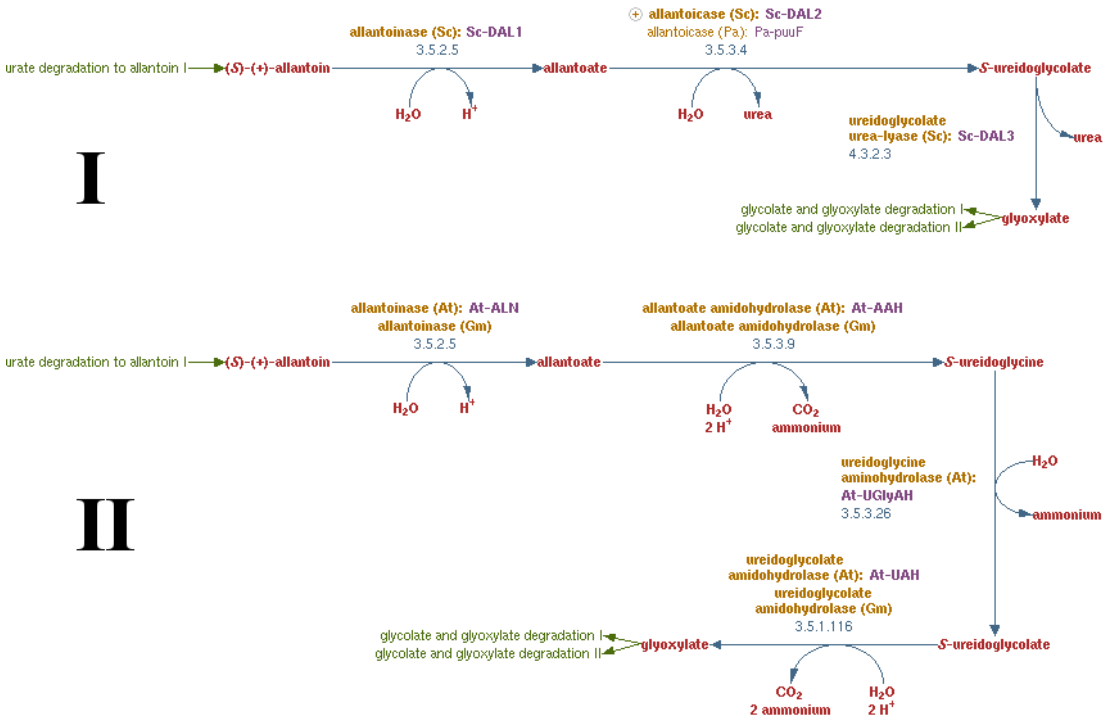

Figure 10: The pathways "allantoin degradation to glyoxylate" I and II from *Saccharomyces cerevisiae* and *Arabidopsis thaliana*, respectively. Both pathways begin and end with the same metabolites, but both take a different route.

terms of their size, but were otherwise selected at random from a set of pathways containing alternate branches.

17.6.1 Reaction correctness

Two factors were used to assess the correctness of substrate-product binary interactions: the presence of the interaction in BRENDA [138][9] and the number of times the interaction was extracted. In total, across the three pathways, 1459 reactions were extracted, of which 220 contained an interaction found in BRENDA. 90% (197) of these reactions were found to be correct extractions accurately reflecting the content of the sentence. The remaining 10% are accounted for by incorrect extractions that correspond to real reactions purely by coincidence.

Of the 1239 reactions not found in BRENDA, 72% (893) were found to be unambiguously incorrect extractions. The remaining 28% were not found in BRENDA for a number of reasons:

- Reactions involving generic molecules (such as *alcohol*) — while BRENDA contains generic reactions, generic molecules cannot be assigned an InChI and, therefore, cannot be cross-referenced with the database.
- Composite reactions — multiple reactions are often described as though they are a single reaction. For instance, a sentence may describe the conversion of A to D without mentioning the intermediate molecules B and C. While composite reactions will not be found in BRENDA the extractions were an accurate representation of the information in the sentence.
- There are a number of possible points of failure in assigning a molecule an InChI, any of which can result in no InChI, or an incorrect InChI, being assigned and preventing cross-referencing with BRENDA. There were cases of metabolites in ChEBI not having comprehensive synonym lists, errors in accurately extracting text from PDFs and bugs in searching the offline ChEBI database.

Of the 1211 reactions that were only extracted once, only 32% (389) were correct extractions. The majority of extractions extracted more than once were correct, however: 60% of those extracted twice, 82% of those extracted three times, 86% of those extracted four times and 95% of

[9] As BRENDA is not freely downloadable, the companion database BKM-React [18] of cross-referenced species non-specific metabolic reactions from BRENDA, KEGG and MetaCyc was used as a substitute.

Factor	Estimated probability of correct reaction
Found in BRENDA	
True	0.90
False	0.28
Number of extractions	
1	0.32
2	0.60
3	0.82
4	0.86
5 or more	0.95

Table 10: A table showing the probabilities of correctness factors derived from the development set of three pathways.

those extracted more than five times. The extraction of the same non-existent reaction multiple times is coincidental and generally involve metabolites found in many pathways (such as *acetyl-CoA*).

Table 10 summarises the probabilities of the two correctness factors. These factors are used to calculate extraction scores for each binary interaction in each extracted reaction. For example, the extraction score for a binary interaction found in BRENDA and extracted only once would be calculated like so:

$$0.9 + (1 - 0.9) \times 0.32 = \mathbf{0.93}$$

Likewise, the extraction score for a binary interaction not found in BRENDA, but extracted four times would be calculated like so:

$$0.28 + (1 - 0.28) \times 0.86 = \mathbf{0.90}$$

17.6.2 Reaction relevance

Reaction relevance is subjective and depends on the needs of the user. For the purposes of training LiMPET and the subsequent test of performance, a binary interaction was defined as relevant if it belonged to the pathway that was being looked for. For instance, when using "allantoin degradation to glyoxylate I" as a seed pathway to extract "allantoin degradation to glyoxylate II" in *Arabidopsis thaliana* from the literature, only those binary interactions found in pathway II were deemed relevant.

As with correctness, a number of potential relevance factors for a binary interaction were identified:

- The presence of the binary interaction in the seed pathway.
- The number of times the reaction is extracted.
- The relevance of the source document(s) — measured by the similarity between the set of metabolites mentioned in the source document(s) and those occurring in the seed pathway.
- The presence of the binary interaction in a branch connecting two metabolites occurring in the seed pathway.

A substrate-product binary interaction found in the seed pathway is given a base score of 1.0 (which results in the final relevance score equalling 1.0), while a base score of 0 is given if not found in the seed pathway.

Across the development set of pathways, relevance was partially correlated with the number of times a reaction was extracted with 18% of reactions extracted five or more times being relevant, 10% of those extracted 4 times and 2% of those extracted 3 times.

To assess the relevance of a specific article, Dice's coefficient [139] was employed. The measure is used to compare the similarity of two samples — in this case, the compared samples are the small molecules mentioned in the article and the metabolites present in the seed pathway. Dice's coefficient is calculated using the same formula as the F-score (see Section 8). For instance, if a seed pathway contains 10 metabolites and an article mentions 5 of these metabolites in addition to a further 20 small molecules not found in the seed pathway, the Dice's coefficient would be calculated as follows:

$$
\begin{aligned}
Precision &= \frac{true\ positives}{true\ positives + false\ positives} \\
&= \frac{5}{5+20} \\
&= 0.20
\end{aligned}
$$

$$
\begin{aligned}
Recall &= \frac{true\ positives}{true\ positives + false\ negatives} \\
&= \frac{5}{5+5} \\
&= 0.50
\end{aligned}
$$

$$
\begin{aligned}
Dice's\ coefficient &= 2 \times \frac{|A \bigcap B|}{|A| + |B|} \\
&= 2 \times \frac{0.20 \times 0.50}{0.20 + 0.50} \\
&= 0.29
\end{aligned}
$$

For each extracted reaction, Dice's coefficient was calculated for each source document. The reactions extracted for each of the three pathways were merged into a single list and ranked by the greatest Dice's coefficient for each reaction. For each reaction the proportion of relevant reactions (i.e. the reactions in the known pathways) was calculated in a window of 50 reactions either side of the reaction. Figure 11 shows a chart plotting the Dice's coefficient against the proportion of relevant reactions. A logarithmic curve with the following equation was fitted to the data[10] to calculate a *document relevance factor*:

$$y = 0.0332686 \ln(x) + 0.113365$$

[10]The regression calculation was performed using Gnumeric, an open-source spreadsheet program for Linux operating systems.

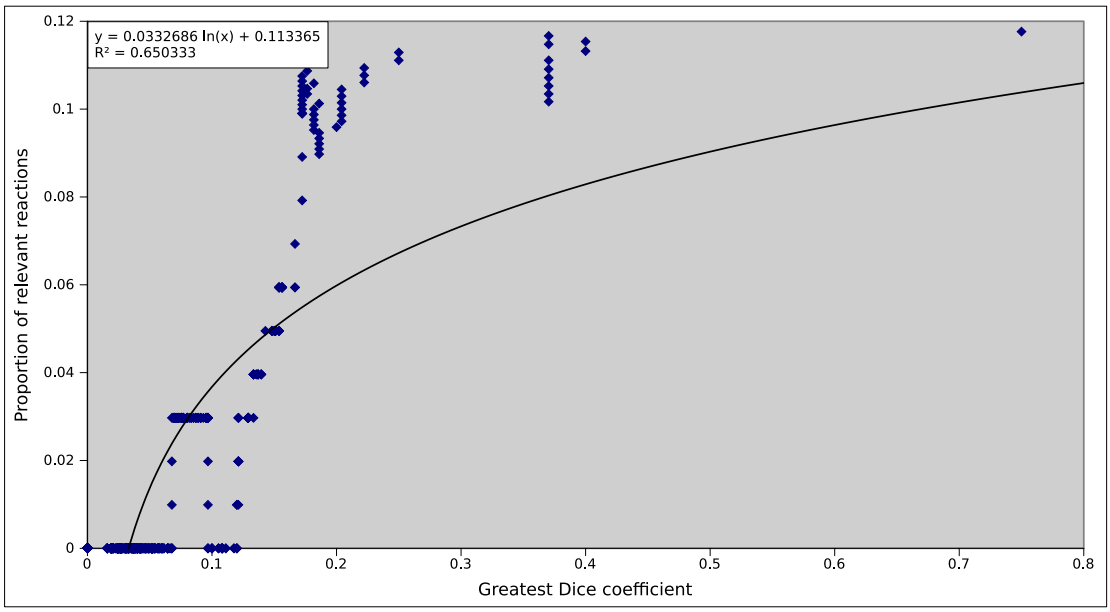

Figure 11: A graph showing the correlation between the greatest Dice coefficient for a set of source articles and the proportion of relevant reactions.

---

where y = document relevance factor, x = greatest Dice's coefficient

Therefore, for a reaction where the greatest Dice's coefficient achieved by a source document is 0.2, the document relevance factor would be calculated like so:

$$y = 0.0332686 \ln(\mathbf{0.2}) + 0.113365$$

$$y = 0.0598213$$

The final factor taken into account in calculating relevance is the presence of a binary interaction in a branch connecting two metabolites found in the seed pathway. Not all branches are of equal quality, however. Longer branches are typically less relevant as they are more likely to be the integration of a separate pathway and not a variation on the seed pathway and branches containing a low confidence reaction are less likely to be real branches. The quality of the branch is calculated by simply multiplying the extraction scores of each reaction in the branch. Therefore, short branches containing reactions with good extraction scores will achieve a high score. This does, however, require all reactions in a branch to be correctly extracted — a single missing reaction will cause the two seed metabolites to no longer be connected.

Consider an extracted pair which is not found in the seed, but is extracted fours times, the greatest Dice's coefficient for any source document is 0.3 and it belongs to a branch containing three reactions connecting two metabolites from the seed pathway (with individual extraction scores for the reactions in the branch of 0.9, 0.8 and 0.7). The relevance score would be calculated as follows:

- Binary interaction not found in seed pathway:

$$x = 0$$

- Reaction is extracted four times, producing a score of 0.10 (as 10% of reactions extracted four times in the development set were relevant):

$$\begin{aligned} x &= 0 + (1 - 0) \times \mathbf{0.10} \\ &= 0.10 \end{aligned}$$

- The greatest Dice's coefficient of any source document is 0.3:

$$\begin{aligned} x &= 0.10 + (1 - 0.10) \times 0.0332686 \ln(\mathbf{0.3}) + 0.113365 \\ &= 0.10 + (1 - 0.10) \times 0.07331051 \\ &= 0.166 \end{aligned}$$

- The binary interaction is part of a branch containing 3 reactions:

$$\begin{aligned} x &= 0.166 + (1 - 0.166) \times (\mathbf{0.9} \times \mathbf{0.8} \times \mathbf{0.7}) \\ &= 0.166 + (1 - 0.166) \times 0.504 \\ &= 0.586 \end{aligned}$$

The calculated relevance score for the pair is 0.586.

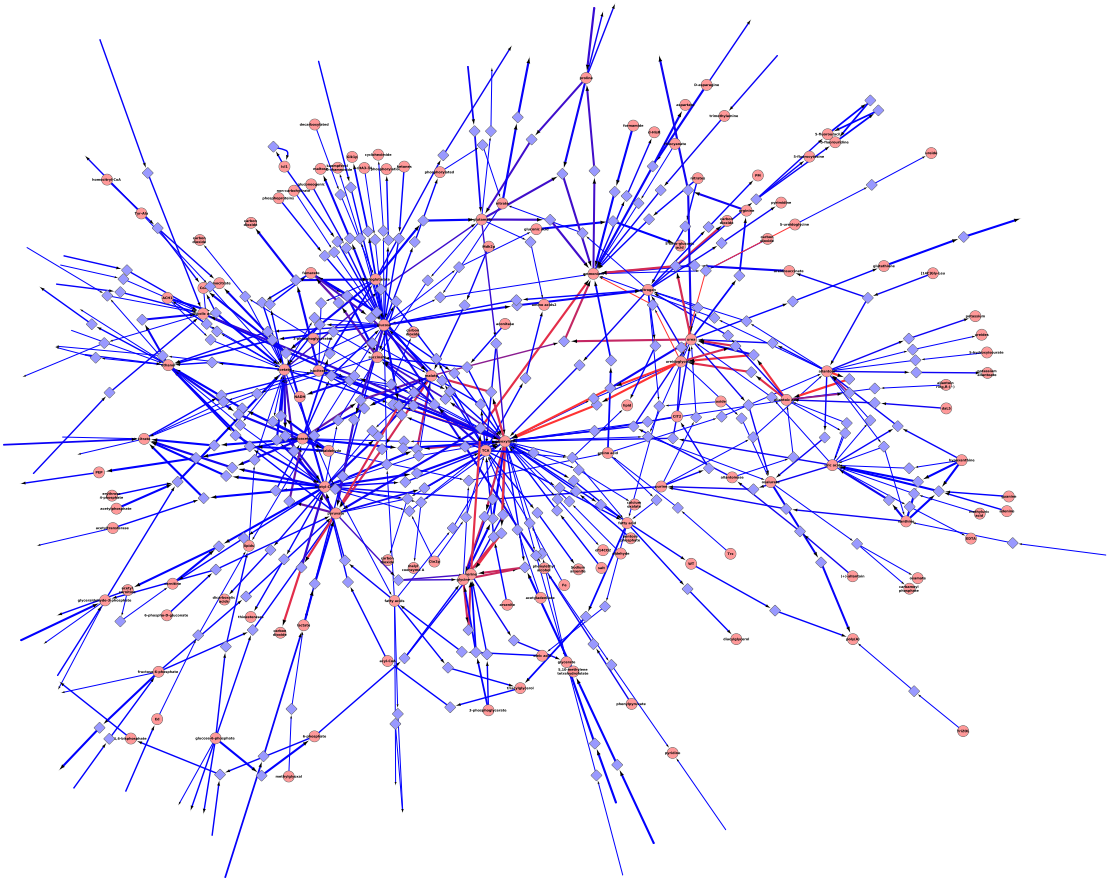

Figure 12: A partial view of a network ("allantoin degradation to glyoxylate" in *Saccharomyces cerevisiae*) extracted by LiMPET. The full pathway is significantly larger and viewing details would not be possible if shrunk to a single page. Metabolites are displayed as pink circles and reaction nodes as blue squares. Extraction scores are proportional to the thickness of the connecting arrows, while relevance scores are reflected by the colour (from blue: low relevance, to red: high relevance. Figure 13 shows the network with extraction and relevance score thresholds applied.

17.7 Program output

Extracted reactions are outputted in SBML — a standard XML-based format designed for the exchange of metabolic pathways [116]. Custom annotations containing the links to source articles and the extraction and relevance scores are assigned to reactions.

While Cytoscape, the standard biological network visualisation tool, can display SBML files, custom annotations cannot be read. LiMPET can, therefore, also output delimited files suitable for visualisation with Cytoscape. Visual properties of the network can then be used to show the extraction and relevance scores of individual reactions (see Figures 12 and 13).

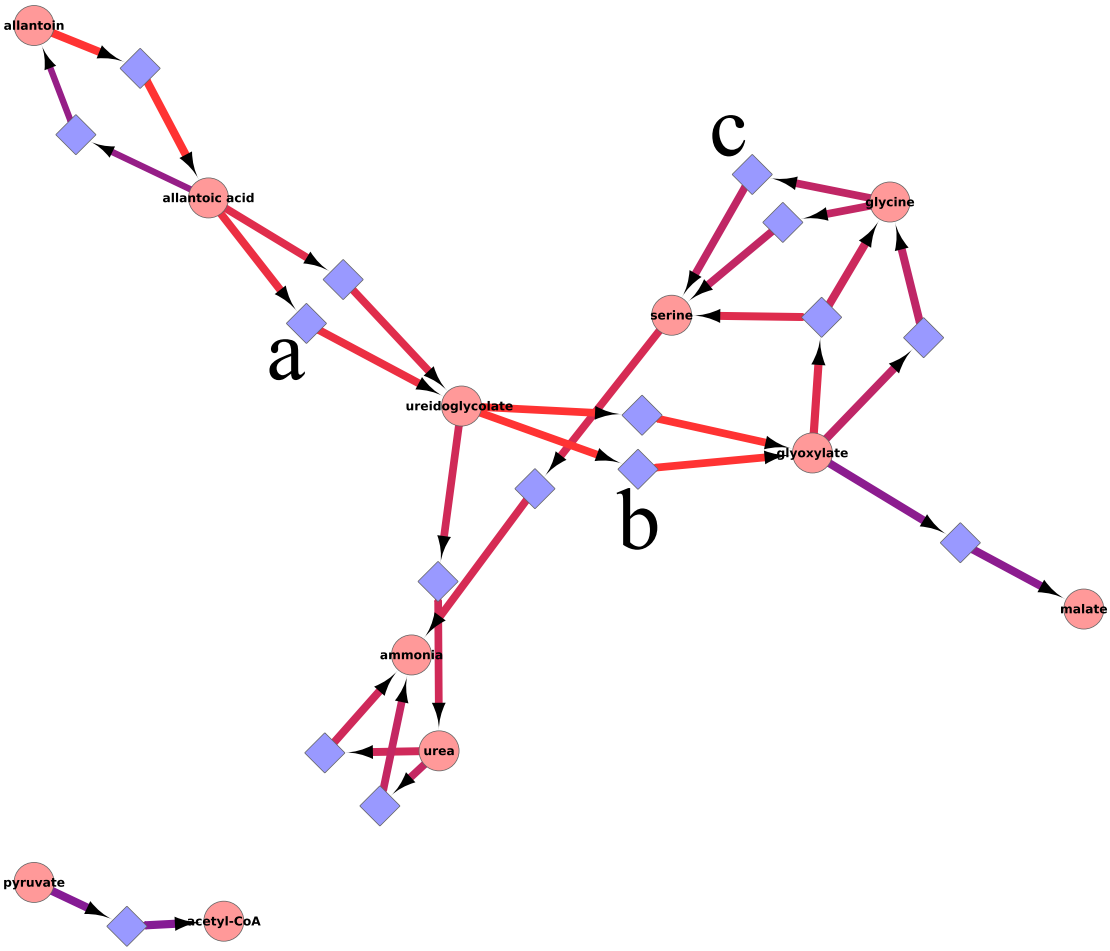

Figure 13: The network shown in Figure 12, but with extraction and relevance score thresholds applied.
Note the seemingly duplicate pairs of reactions joining metabolites (a, b and c). These pairs of reactions consist of one reaction containing the side metabolites and one not containing them (this shows the under-merging described in Section 17.5). The side metabolites have been assigned relevance scores below the threshold and, therefore, cannot be seen in this network.

17.8 Evaluating LiMPET

In addition to the development pathways, a further 14 MetaCyc pathways were identified with an alternative pathway describing a different route between two metabolites. Branches varied in length between 1 and 3 reactions. While there are more examples of alternative branches (many longer in length) in MetaCyc, it was necessary that the pathway was fully annotated in at least one species. Unfortunately many pathways in MetaCyc are formed from the merging of incomplete pathways from multiple species; a drawback which is more prevalent with longer pathways.

An extraction cut-off was applied to all extracted reactions (such that reactions not found in BRENDA and only extracted once were below the threshold) and reactions were ranked by relevance score. TAP-*k* was used to assess LiMPET's performance (see Section 8.1.3 for a detailed discussion of performance metrics and why TAP-*k* was chosen).

18 Results

In total, across all 14 pathways, over 4 500 articles were retrieved and mined. LiMPET succeeding in extracting and scoring highly all reactions in 9 of the pathways, while no reactions were retrieved in 3 of the pathways. For the purposes of evaluation, reactions were assigned the highest extraction and relevance score achieved by their individual metabolites. An extraction threshold of 0.75 was set and the remaining reactions were ordered by relevance score. Table 11 shows the rankings of extracted reactions corresponding to reactions in the pathway being reconstructed (see Appendix V for an example ranked list) and TAP-*k* scores for all 14 pathways. Appendix VI shows a selection of extracted pathways compared to their MetaCyc counterparts.

TAP-*k* scores showed significant variation, even amongst completely extracted pathways. The highest TAP-*k* scores were achieved by the extraction of "indole-3-acetate biosynthesis III" in *Pseudomonas savastanoi*, while the extraction of "indole-3-acetate biosynthesis I" in *Arabidopsis thaliana* achieved far lower scores, despite both pathways being extracted correctly. This is because the short length of the pathways has caused the TAP-*k* measure to become particularly sensitive to a relatively small number of highly ranked irrelevant reactions.

Seed pathway	Attempted reconstruction	Organism	Ranking of relevant reactions	Mean EPQ (k)		
				5	10	20
allantoin degradation to glyoxylate II	allantoin degradation to glyoxylate I	*Saccharomyces cerevisiae*	1, 3, 4	0.6300	0.6269	0.6181
glycerol degradation II	glycerol degradation I	*Arabidopsis thaliana*	1, 2	0.7778	0.7222	0.7037
glycerol degradation I	glycerol degradation II	*Klebsiella pneumoniae*	2, 3	0.4841	0.4402	0.4127
indole-3-acetate biosynthesis III (bacteria)	indole-3-acetate biosynthesis I	*Arabidopsis thaliana*	5, 6	0.2068	0.1993	0.1929
indole-3-acetate biosynthesis I	indole-3-acetate biosynthesis III (bacteria)	*Pseudomonas savastanoi*	1, 2	0.7778	0.7778	0.7619
ent -kaurene biosynthesis II	ent-kaurene biosynthesis I	*Arabidopsis thaliana*	3, 4	0.4111	0.3519	0.3081
ent-kaurene biosynthesis I	ent -kaurene biosynthesis II	*Physcomitrella patens*	4	0.2250	0.2083	0.1513
mannosylglycerate biosynthesis II	mannosylglycerate biosynthesis I	*Pyrococcus horikoshii*		0	0	0
mannosylglycerate biosynthesis I	mannosylglycerate biosynthesis II	*Rhodothermus marinus*	3	0.3333	0.2917	0.2292
methylglyoxal degradation V	methylglyoxal degradation VII	*Pseudomonas putida*		0	0	0
phenylalanine biosynthesis II	phenylalanine biosynthesis I	*Bacillus subtilis*	1, 31	0.2708	0.2625	0.2796
phenylalanine biosynthesis I	phenylalanine biosynthesis II	*Nicotiana sylvestris*		0	0	0
pyruvate fermentation to ethanol II	pyruvate fermentation to ethanol I	*Chlamydomonas reinhardtii*	6, 7, 25	0.1658	0.1623	0.1594
pyruvate fermentation to ethanol I	pyruvate fermentation to ethanol II	*Zea mays*	2, 3	0.4444	0.4259	0.4156
			Average	0.3376	0.3192	0.3023

Table 11: The ranking of relevant reactions and the TAP-k scores achieved by LiMPET in the attempted reconstruction of 14 MetaCyc pathways. In the ranking column green signifies a complete extraction; orange, a partial extraction; red, a failed extraction.

18.1 Error analysis

There were three categories of error found: failures to retrieve articles describing the relevant reactions, failures to extract the relevant reactions from articles that have been retrieved and failures to assign extracted reactions with a high relevance ranking.

Failure to retrieve the appropriate articles had a number of different causes:

- PubMed indexes the exact spelling in article records. While LiMPET's search strategy involves including organism and small molecule synonyms in the search query and PubMed automatically includes synonyms known synonyms, the use of non-standard terms or misspellings by authors can result in a failure to retrieve articles. The attempted reconstruction of "phenylalanine biosynthesis II" in *Nicotiana sylvestris* failed to extract any relevant reactions. This is due to most work in this area being carried out by a single group who prefer the spelling *Nicotiana s**i**lvestris* — a variant spelling not present in NCBI Taxonomy. If the variant spelling is included in the LINNAEUS dictionary, the articles are successfully retrieved and the relevant reactions extracted.

- By default LiMPET attempts to retrieve the 100 most recent articles for a given article. One citation for "phenylalanine biosynthesis I" in *Bacillus subtilis* in MetaCyc, was found by the search, but fell outside the 100 most recent articles. The article [140], published in 1976, was the 275^{th} most recent article in the returned set from one query.

- PubMed principally includes records for research articles, whereas reactions in MetaCyc are obtained from additional sources, such as books. Evidence for one reaction in "phenylalanine biosynthesis I" in *Bacillus subtilis* was obtained from a book.

- The single reaction MetaCyc pathway "methylglyoxal degradation VII" in *Pseudomonas putida* originates from a single source article. The article was published in the now defunct journal *Agricultural and Biological Chemistry*, which is not indexed in PubMed and so could not be retrieved by LiMPET. The article could be found using a manual Google Scholar search, however.

With successfully retrieved articles, there are two principal causes of errors:

- Following the extraction of a putative reaction, failure to assign an InChI for a component metabolite effectively prevents the reaction from achieving high extraction and relevance

scores as the reaction cannot be combined with or linked to other reactions and cannot be cross-referenced with BRENDA. In the extraction of "mannosylglycerate biosynthesis I" in *Pyrococcus horikoshii*, LiMPET successfully retrieved the two adjoining reactions *GDP-mannose* to *mannosyl-3-phosphoglycerate* to *mannosylglycerate*, but the extraction score for both fell below the cut-off. This was due to a failure to cross-reference *mannosyl-3-phosphoglycerate* with ChEBI, which listed two more complex synonyms for the metabolite: *2-(α-D-mannosyl)-3-phosphonatoglycerate(3–)* and *2-(α-D-mannosyl)-3-phosphonatoglycerate trianion*.

- As discussed in Section 15, LiMPET's pattern-based core algorithm is unable to extract reactions from descriptions that do not match one of the manually defined patterns. For instance, in the extraction of "phenylalanine biosynthesis I" in *Bacillus subtilis*, LiMPET was unable to extract the reaction *phenylpyruvate* to *phenylalanine* from the following sentence:
  This reaction is a transamination step involving glutamate and either p-hydroxyphenylpyruvate, the precursor to tyrosine or phenylpyruvate, the precursor to phenylanine.

Failures to correctly assess relevance were also identified:

- As was discussed previously, a single reaction in the pathway "phenylalanine biosynthesis I" in *Bacillus subtilis* was missed entirely. As this reaction is part of an alternative branch containing two reactions, the other reaction, which was correctly extracted was assigned a low relevance score as the branch could not be completed.
- In "pyruvate fermentation to ethanol II" in *Chlamydomonas reinhardtii*, one reaction in the "novel" branch could not be cross-referenced with BRENDA, resulting in mediocre extraction and branch scores. Compounding the error, the pathway involves metabolites involved in many cellular pathways (such as pyruvate, acetyl-CoA and acetaldehyde) which resulted in a large number of high scoring branches.

18.2 Extracting pathways from abstracts and PMC-OA full-text articles

In section 17.2 I described the method used for obtaining full-text articles by screen-scraping publishers' websites. As this behaviour cannot be released in the final tool, however, the evaluation of LiMPET was repeated, but the retrieval of text was limited to abstracts and open-access

PMC articles that could be retrieved using the NCBI E-Utils API. Table 12 shows a comparison of the recall achieved compared to the original performance using screen-scraping to retrieve all available full-text articles. For the purposes of this analysis, a reaction was considered to be successfully retrieved regardless of its relevance ranking to better determine how much available data is in the open-access literature. Despite these lenient conditions the limitation in text to mine produced a significant drop in performance with only a single pathway being completely extracted and no relevant reactions being extracted from 8 pathways.

19 Discussion

Here I have described the expansion of LiMPET from a core metabolic reaction extraction algorithm to a pipeline able to retrieve relevant articles and to construct individually extracted reactions into pathways. LiMPET performed well with its ability to retrieve novel alternative branches for a given pathway in specific organisms, with 9 of the 14 evaluation pathways being retrieved completely with individual reactions scored appropriately. While the TAP-*k* scores achieved are largely in line with those from the gene normalisation domain [45, 141, 142], they are difficult to interpret in this context due to the low number of possible relevant extractions. Despite this, a subjective analysis of the rankings of individual relevant reactions show a good performance with relevant reactions typically being placed within the top ten.

The partial and failed pathway extractions highlight weaknesses of the method and tool. While certain errors, such as defunct journals not being indexed by PubMed, cannot be solved by improving the method, potential improvements could be made to the search strategy and the core algorithm. The significant drop in performance when the evaluation was repeated using just abstracts and full-text articles from the PMC-OA Subset, however, raises concerns about whether a substantial performance increase could be found without access to more full-text articles. The implications of these findings are discussed in Part VI.

Attempted reconstruction	Organism	Total reactions in pathway	Recall from all full-text	Recall from abstracts and PMC-OA articles
allantoin degradation to glyoxylate I	*Saccharomyces cerevisiae*	3	1.00	0.33
glycerol degradation I	*Arabidopsis thaliana*	2	1.00	0.00
glycerol degradation II	*Klebsiella pneumoniae*	2	1.00	0.00
indole-3-acetate biosynthesis I	*Arabidopsis thaliana*	2	1.00	0.50
indole-3-acetate biosynthesis III (bacteria)	*Pseudomonas savastanoi*	2	1.00	0.50
ent-kaurene biosynthesis I	*Arabidopsis thaliana*	2	1.00	0.00
ent -kaurene biosynthesis II	*Physcomitrella patens*	1	1.00	1.00
mannosylglycerate biosynthesis I	*Pyrococcus horikoshii*	2	0.00	0.00
mannosylglycerate biosynthesis II	*Rhodothermus marinus*	1	1.00	0.00
methylglyoxal degradation VII	*Pseudomonas putida*	1	0.00	0.00
phenylalanine biosynthesis I	*Bacillus subtilis*	3	0.67	0.33
phenylalanine biosynthesis II	*Nicotiana sylvestris*	3	0.00	0.00
pyruvate fermentation to ethanol II	*Chlamydomonas reinhardtii*	2	1.00	0.00
pyruvate fermentation to ethanol I	*Zea mays*	3	1.00	0.33

Table 12: A comparison of the recall achieved by LiMPET when extracting pathways from all available full-text retrieved by screen-scraping with the extraction of pathways from just abstracts and PMC-OA articles.

Part V

Towards the automated annotation of BioCyc predicted pathways

20 Introduction

The use of text-mining is increasingly common in the curation of gene information for model organism databases. The maintainers of WormBase [143], a gene-centric database for the model organism *Caenorhabditis elegans*, incorporate text-mining methods into their curation workflow [144]. New articles are flagged as containing certain data types (such as expression patterns and genetic interactions) and named entity recognition (of entities such as transgenes and molecules) is carried out using the Textpresso information retrieval system [145]. The WormBase team have collaborated with the maintainers of the model organism databases dictyBase (*Dictyostelium discoideum*) and TAIR (*Arabidopsis thaliana*), to implement their pipeline in the curation of gene data from other organisms. The FlyBase Consortium are also actively investigating the implemenation of text-mining methods in their curation workflow [8].

BRENDA [5] was originally a purely manually annotated database, but has since supplemented this manually extracted data with data retrieved using text-mining methods. Unlike the biocuration of gene-centric model organism databases, however, data from text-mining is used to supplement manual curations (and is clearly marked as such). The subset FRENDA (Full Reference ENzyme DAta) contains data obtained by mining article titles and abstracts for co-occurring enzyme and organism names, while AMENDA (Automatic Mining of ENzyme DAta) uses ontologies to retrieve the tissue source and subcellular localisation of enzymes. While it would require significant man-power to manually determine all the organisms that a particular enzyme is known to be present in, FRENDA is able to provide relevant references to the user who can then rapidly check the extracted information.

I believe that metabolic pathway databases, such as BioCyc, could greatly benefit from the addition of data retrieved using text-mining similar to BRENDA. Here an example use-case will be shown: using LiMPET to identify evidence in the literature for predicted pathways in

BioCyc.

As of August 2014, MetaCyc contains 2 569 annotated pathways and BioCyc contains 3 563 organism databases containing predicted pathways. While most MetaCyc pathways will not have a corresponding predicted pathway in every organism database, there is still too much available data to attempt to corroborate every predicted pathway for every organism. LiMPET's extraction process involves two long-running sub-tasks: downloading full-text articles and extracting reactions from the text. The MetaCyc pathways used to assess LiMPET's performance in Part IV were modest in size, resulting in no more than 500 papers being downloaded and mined for a single test pathway — a maximum of 3 hours was taken for a single pathway extraction. As we investigate larger pathways and using multiple variant pathways as seeds, however, the run time will increase dramatically[11]. A set of pathways in a single organism were, therefore, selected for investigation.

21 Pathway and organism selection

Previously I have described attempted pathway extractions in many organisms ranging from thoroughly studied model organisms, such as *Arabidopsis thaliana*, to organisms, such as *Nicotiana sylvestris*, with relatively few citations. Model organisms are unsuitable for this task as there are specialist teams that curate relevant research for the Tier 1. While retrieving reactions present in organisms with few citations proved successful in a number of cases (see Section 18), it was known beforehand that the reactions were present in the literature. For a lesser known organism, such as *N. sylvestris*, where only 372 article records in PubMed reference the organism, few pathways will have been characterised. The ideal organism for this task is one that is well studied, but has few curated pathways in MetaCyc. *Mycobacterium tuberculosis*, with its effect on human health, is well-studied with approximately 50 000 potentially relevant articles in PubMed, but only 46 pathways in BioCyc are listed as being experimentally observed in the organism.

The identity of these 46 pathways is perhaps evidence of bias towards those of therapeutic interest (such as mycothiol biosynthesis), but whether that is a bias in the curation of MetaCyc

[11] Abstracts and PMC open-access full-text articles, however, can be retrieved rapidly as the NCBI E-Utils API allows bulk downloading and no screen-scraping is required. Attempts were made to parallelise the core algorithm, but the architecture of certain third party libraries prevented this. It would be possible to run algorithm is separate processes (in a cluster, for instance), but the memory needs would be considerable.

pathways or in the focus of original research is unclear. A further 1 117 pathways are predicted to to be present, however. Of these pathways I chose a subset of "core" pathways (the 20 amino acid biosynthesis pathways) that I considered to be likely candidates for research and, therefore, more likely than a purely random selection to have supportive evidence in the literature that has not been curated by BioCyc.

For each amino acid, LiMPET was provided with each of its biosynthesis pathways in MetaCyc as seed pathways. For instance, in the case of asparagine biosynthesis, the three pathways "asparagine biosynthesis" I, II and III were merged and used as a seed pathway despite pathway I being the only pathway predicted to exist in *M. tuberculosis*. All variants were used as this would increase the likelihood of finding variants with evidence in the literature that have not been predicted to be present by BioCyc. For each PubMed query the most recent 100 articles were downloaded — the total number of articles to mine for a single pathway ranged from 362 to 937 articles.

22 Results

Table 13 shows the amino acid biosynthesis pathway variants predicted to be present in *M. tuberculosis* and the variants for which evidence LiMPET was able to extract from the literature.

Table 13: A table showing the amino acid biosynthesis pathway variants predicted to be present in *M. tuberculosis* by BioCyc and the coverage of each pathway by reactions extracted from the literature using LiMPET. The cell colour indicates the level of coverage, with green showing a complete pathway extraction, orange showing a partial extraction and red indicating that no reactions in the pathway were extracted. Blue cells show extracted pathways that were not predicted to be present by BioCyc.

Amino acid biosynthesis pathway	Variants predicted in M. tuberculosis in BioCyc	Pathway coverage by LiMPET
Alanine	I	0/3
	II	1/1
	III	1/1
Arginine	I	0/9
	II	0/8
Asparagine	I	0/1
Aspartate	I	0/1
	II	0/2
Cysteine	I	2/2
	II	0/2
Glutamate	I	1/1
Glutamine	I	1/1
Glycine	I	0/1
	III[12]	1/1
Histidine	I	0/10
Isoleucine	I	2/13
	I (from threonine)	0/7

[12]Glycine biosynthesis III was not predicted to be present by BioCyc, but was extracted by LiMPET.

Table 13: A table showing the amino acid biosynthesis pathway variants predicted to be present in *M. tuberculosis* by BioCyc and the coverage of each pathway by reactions extracted from the literature using LiMPET. The cell colour indicates the level of coverage, with green showing a complete pathway extraction, orange showing a partial extraction and red indicating that no reactions in the pathway were extracted. Blue cells show extracted pathways that were not predicted to be present by BioCyc.

Amino acid biosynthesis pathway	Variants predicted in M. tuberculosis in BioCyc	Pathway coverage by LiMPET
Leucine	I	0/6
Lysine	I	2/9
	VI	0/7
Methionine	I	0/5
	II	0/6
	III	0/2
Phenylalanine	I	3/3
Proline	I	0/4
Serine	I	3/3
Threonine	I	0/6
	I (from homoserine)	0/2
Tryptophan	I	3/6
Tyrosine	I	3/3
Valine	I	2/4

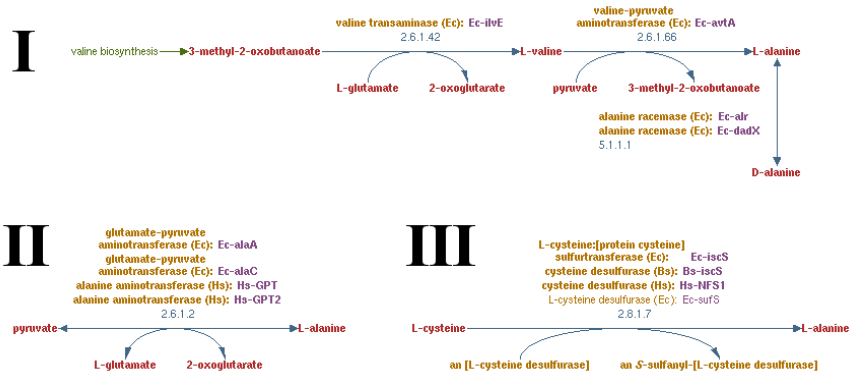

Figure 14: The three alanine biosynthesis pathways in MetaCyc.

LiMPET was able to corroborate a number of BioCyc predicted pathways. Of the 20 amino acids, at least one complete biosynthesis pathway corresponding to a MetaCyc pathway was extracted for 7 amino acids. Incomplete biosynthesis pathways were extracted for a further 4 amino acids. While no reactions from the biosynthesis pathways for the other 9 amino acids were extracted, this is not necessarily a fault of LiMPET — the pathways may not have been characterised in the literature.

For instance, consider alanine biosynthesis. Figure 14 shows the three experimentally verified pathways in MetaCyc — all of which are predicted to exist in *Mycobacterium tuberculosis*. LiMPET was able to extract both the forward and backward reactions of pathway II and the reaction of pathway III, but was unable to find either of the pathway I reactions. Consider the reaction pyruvate to L-alanine. LiMPET returned links to the two source papers and the specific sentences that this reaction was extracted from:

> "L-AlaDH catalyzes the NADH-dependent reversible oxidative deamination of l-alanine to pyruvate and ammonia." [146]
>
> "NAD(H)-dependent l-AlaDH catalyze the oxidative deamination of l-alanine to pyruvate and ammonia (catabolic reaction) or, in the reverse direction, the reductive amination of pyruvate to l-alanine (biosynthetic reaction)."[13] [147]

[13]Note that the erroneous hyphen in the final word *reaction*, is due to the sentence being extracted from a PDF, where there is no differentiation between soft and hard hyphens. While in this case the soft hyphen could be eliminated using an English dictionary, differentiating soft and hard hyphens in specialist terms (such as small molecule names) is a more difficult task.

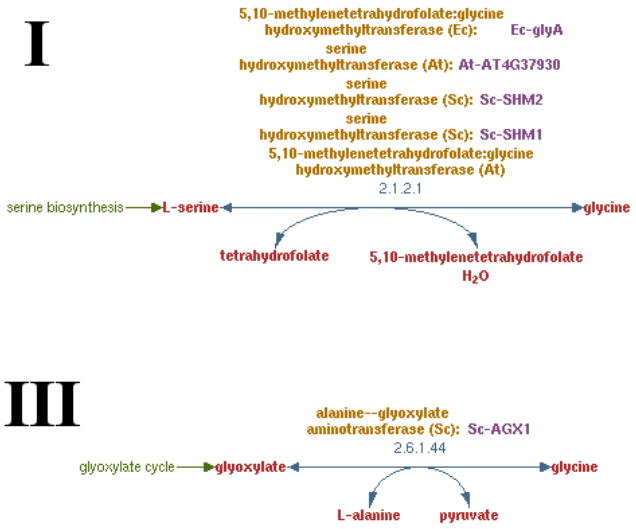

Figure 15: The pathways "glycine biosynthesis I" and "glycine biosynthesis III" from MetaCyc.

In the case of glycine biosynthesis, LiMPET was able to extract information that contradicted the BioCyc prediction. Figure 15 shows two alternate glycine biosynthesis pathways from MetaCyc. While BioCyc only predicted the presence of "glycine biosynthesis I" in *M. tuberculosis*, LiMPET was able to extract "glycine biosynthesis III", while finding no evidence for the predicted pathway.

The glyoxylate to glycine reaction was extracted from three different sentences in a single article [148]:

> "The putative glycine dehydrogenase of *Mycobacterium tuberculosis* catalyzes the reductive amination of glyoxylate to glycine but not the reverse reaction."
>
> "GxRA is involved in the reductive amination of glyoxylate to glycine."
>
> "This enzyme was detected by the reductive amination of glyoxylate to glycine concurrent with the oxidation of NADH to NAD+ (Fig. 1)."

Moreover, a manual search for the predicted pathway using PubMed, Google Scholar and Google was not able to find any evidence of the reaction's presence in *M. tuberculosis*. While this does not mean that the BioCyc prediction is incorrect, the failure to predict the presence of the pathway "glycine biosynthesis III" certainly is incorrect. In the description of the "glycine

biosynthesis III" pathway it is noted that the pathway has been documented in archaea, bacteria and eukaryotes, but it is unclear why it was not predicted for *M. tuberculosis* specifically.

A number of pathways were partially extracted. In the case of tryptophan and valine biosynthesis, LiMPET was able to extract half of the reactions from the predicted pathways.

23 Discussion

Here I have shown an example use-case for LiMPET: the automated extraction of literature evidence for predicted pathways. Using collections of MetaCyc pathways I was able to use LiMPET to find corroborating evidence for a number of amino acid biosynthesis pathways in *Mycobacterium tuberculosis* and contradictory evidence for one pathway.

BioCyc tier 1 databases have been used to assess the performance of LiMPET previously — namely EcoCyc (see Part III) and MetaCyc (see Part IV). Using these tier 1 databases, we can be confident that descriptions of the curated metabolic reactions exist in the literature. Therefore, whenever LiMPET fails to extract a pathway or a particular reaction from the literature, we can be confident that some part of the methodology has failed and these failures can be further investigated. In real use and in this performance assessment, however, there is less certainty that the information is even in the literature.

Using the functionality shown here, LiMPET could be used to help database curators and users, alike. Using LiMPET, I was able to extract evidence for seven complete amino acid biosynthesis pathways. The only manual step needed for each pathway was the retrieval of MetaCyc IDs for the seed pathway; a process requiring only a few minutes in each case (as described in Section 17, the BioCyc API cannot be used to search the database by pathway name — instead the pathway IDs must be manually extracted from the webpage URLs). A curator would potentially only need to read a select few papers in depth to ensure extractions were correct.

Currently BioCyc predicted pathways can provide a useful start for a researcher interested in a particular pathway. As the reliability of predictions is not known, however, individuals still need to carry out their own research of the literature. Using LiMPET a single researcher could potentially find evidence for a predicted pathway without needing to manually search the literature.

Perhaps the most compelling potential use, however, is the automated annotation of predicted pathways with evidence from the literature, similar to how BRENDA displays species-specific enzyme data from FRENDA and AMENDA (and clearly labels the origin of the data).

Once the pathway and organism identifiers have been retrieved, the running time of the program varies greatly depending on the size of the pathways and literature coverage of the organism of interest. At the high end, the four glutamine biosynthesis pathways in MetaCyc contain 16 non-currency molecules (see Section 17.5 for a description of currency molecules), resulting in the retrieval of 937 articles. At the low end, the single valine biosynthesis pathway resulted in the retrieval of just 362 articles. The amount of retrieved articles also affects the time taken to extract reactions.

The attempted extraction of the lysine biosynthesis pathway shows the brittleness of some parts of the method. A minireview documenting the entire pathway in *M. tuberculosis* [149] was retrieved and all reactions were successfully extracted. The majority of the reactions were assigned to the incorrect organism, however. This was due to the title of the article, which referred to *M. tuberculosis*, not being extracted with the rest of the article from the article's webpage. Much of the original research referenced by the review was successfully found, but could not be retrieved because they were published by smaller publishers which the article downloading software could not access.

While this work has shown the potential of LiMPET in assisting database curation, it has further shown that the weak link in the process is the retrieval of full-text articles. I will discuss the implications of this in Part VI.

Part VI

Conclusions and further work

In this thesis I have described the development of LiMPET, a tool for the automated extraction of metabolic pathways from research articles. Initial work focused on the development of a prototype text-mining algorithm for the extraction of individual metabolic reactions. As the bioinformatics community has focused on the extraction of other types of interactions and this problem had yet to be tackled, expectations of this relatively simple pattern-based approach were rather modest. When tasked with extracting reactions from articles known to describe various *E. coli* pathways, however, the algorithm outperformed these expectations significantly. While it had initially been planned to either improve the prototype algorithm or replace it entirely with a more sophisticated method at this stage, it was decided to focus instead on other aspects in order to produce a usable tool.

Following the extraction of individual metabolic reactions by the core algorithm, reactions are assigned to a host organism as articles may contain references to multiple organisms. Extracted metabolites are cross-referenced with ChEBI and InChIs assigned to allow the merging of separate mentions of the same reaction and to join reactions together to form pathways. Extracted reactions are then scored on their correctness and relevance to allow the user to find the particular reactions of interest.

While LiMPET has shown good performance, it is important to recognise the tool's weaknesses and how they could be mitigated:

- The development of LiMPET has been limited by the lack of a metabolic reaction corpus. Ideally, text-mining methods are tested by mining a corpus of text which can then be compared to gold-standard annotations with recall, precision and F-score as standard measures with which to report these results. Without a suitable corpus, non-standard testing methods using relatively small datasets have been employed and the results produced cannot be directly compared to other tools[14]. While the manual error analyses employed give good insight into the performance of LiMPET and would allow further

[14]While there are no other metabolic pathway extraction tools to compare with, comparisons with tools in related domains, such as protein-protein interaction extraction would be useful. It is also my hope that other metabolic pathway extraction methods will be developed in the future — a corpus would allow direct comparisons with LiMPET.

development of the method, development of a machine learning method is not possible without a corpus.

- The core algorithm developed here was originally conceived with the baseline PPI extraction algorithm from Kabiljo *et al.* [1] in mind. Despite the relatively simple heuristic methodology, however, the algorithm performed well and LiMPET was built around it. Limitations of the algorithm were recognised, such as difficulties extracting reactions involved in fatty acid biosynthesis. While the previously described lack of a metabolic reaction corpus would hinder the development of a supervised machine learning method, a semi-supervised learning method would potentially be feasible.

  While supervised learning takes advantage of manually labeled text to train a statistical method, this labeling is an expensive process. Unsupervised learning methods, however, are able to extract string of words between entities in unlabeled text and identify patterns which can then be mapped on to specific relationships [150]. As these methods have no way to determine what an interested relationship is, however, these patterns may be difficult to map to relationships in certain knowledge domains.

  Distant supervision, a semi-supervised learning method, can use a heuristic function or knowledge base to weakly label text and identify sentences likely to express a particular relationship [151]. An unsupervised learning method can then be used to identify patterns in the sentences with the knowledge that the identified patterns can likely be mapped to the relationship of interest. This technique could potentially be applied to metabolic reaction extraction using a database of known metabolic reactions, such as BRENDA (where sentences containing all metabolites contained in a single known reaction would be identified as likely containing a reaction description), and would be a worthwhile avenue for further development.

- The algorithm to assign reaction extractions to their host organisms is quite brittle. Similar heuristic methods to the algorithm developed here were used in entries to the gene normalisation task of BioCreative III [45]. The algorithm assumes that articles have a primary organism of interest (which will be the first organism mentioned in the article) and sentences describing reactions belonging to other organisms will also contain the organism name. These assumptions have generally been found to be correct, but articles involving multiple organisms (such as a review article comparing enzymes across a

genus) can easily confuse the algorithm. A failure to identify the first organism named in the article (for instance, if the article is not extracted fully or extraction from a PDF produces some garbled text) can result in an incorrect identification of the article's primary organism. A small corpus of full-text articles with annotated species names and metabolic reactions was created to develop the algorithm, but a larger corpus would allow the development of a more sophisticated statistical method.

- Assigning InChI identifiers to extracted small molecules is a critical step, allowing reactions to be linked together and cross-referenced with BRENDA. This is carried out using an offline ChEBI database with pregenerated name variants. While the variants account for the presence or absence of expected additional elements (such as stereochemical identifiers and hyphens) a simple misspelling or non-standard spelling can cause a search failure. For instance, consider the small molecule *D-glucose 1,6-bisphosphate*. The search strategy would identify *glucose-bisphosphate* as the same molecule, but not *D-glucose 1,6-bisphos**f**ate*, despite the latter's closer spelling overall. While cases of being unable to find correctly named molecules are unusual, their detrimental effect to the rest of the pipeline may warrant a more sophisticated search method, such as a search index capable of fuzzy string matching.

- No attempt is currently made to distinguish the stereochemistry of an extracted molecule. For example, *alanine*, *D-alanine* and *L-alanine* are all assigned the same InChI. This is due to the stereochemistry typically not being included by the author when the specific enantiomer can be easily inferred (e.g. *glucose* typically refers to *D-glucose* as *L-glucose* is rarely found in nature). Unlike most human readers, however, computer programs do not have the benefit of a scientific education and cannot infer such information.

- A reaction's extraction score is determined by its presence in BRENDA and the number of times it is extracted. It does not, however, take into account the sentence content. As such the score is not very granular and novel reactions that are only found once have the potential to be overlooked. Unfortunately the scores produced by the core algorithm can only be used to compare reaction assignments in the same sentence and are not comparable between extractions from separate sentences[15]. Developing a sentence complexity

[15]For instance, an extracted reaction containing many substrates and products from one sentence will typically score higher than an extracted reaction from a different sentence containing just a single substrate and product, regardless

measure that takes into account factors not currently considered (such as the presence of negation words and the length of the sentence) could alleviate this problem.

- The concept of a "relevant" reaction is subjective and is dependent on the needs of the user. While there are a number of different factors taken into account in scoring a reaction's relevance, the method here focuses on finding alternate routes between two metabolites. Different relevance algorithms could be developed. For example, the user could be interested in links between pathways.

- The extraction and relevance scoring methods were refined using a relatively small training set of three pathways. Despite the limited number of pathways, approximately 1 400 putative reactions were extracted in total — all of which were manually assessed. It would be possible to refine the scores using a larger set, but this would require significantly more man-power.

- The branch finding algorithm is brittle. The algorithm relies on a complete, uninterrupted path between two metabolites. If a single reaction is missed, all other reactions in the branch will achieve a low relevance score if the branch is not known in other organisms. It would be possible to identify potential reactions that bridge a gap that exist in other organisms using BRENDA. Unfortunately assessing the plausibility of such links existing would require species-specific information (to identify the closest related organism containing the linking reactions) which would require access to the BRENDA commercial version.

Despite these various improvements that could be made, it is clear that the greatest weakness of LiMPET, and by extension other text-mining tools, is the retrieval of text to mine.

After the development of the core algorithm I identified the automated retrieval of full-text articles as potentially the most important component in developing a usable text-mining tool. The retrieval of full-text articles (from non-PMC sources) was not included in any published text-mining tool and the subject was never broached in any related articles. I saw this as a problem that had a technical solution and I set out to develop it.

I developed a system which followed links in PubMed article records to publishers' websites. Once connected to the publisher's website the system could then follow links based on

of the quality of the extraction, as the scoring algorithm increments the score for each metabolite included (see Section 12.3).

their text (such as *Full Text (HTML)*) to find the full-text article in either HTML of PDF. The method performed well and LiMPET gained the ability to retrieve the full-text of any article. It was shown that LiMPET performs significantly better when it is able to retrieve large quantities of full-text articles compared to when only abstracts and open-access full-text articles are available (see Section 18.2).

In my enthusiasm to solve this problem, however, I failed to see that the problem wasn't a technical one, but was a political and legal one (see Section 5 for a full discussion). While retrieving articles for the work in Part V, one publisher website detected the use of an automated program and blocked its access. It became clear that releasing a tool with such functionality would be irresponsible. The public release of LiMPET, therefore, only has the ability to automatically download article abstracts and full-text articles from the PMC Open-Access Subset (although locally stored articles can also be mined).

Although I identified several potential algorithmic improvements above, I am skeptical about how large an improvement in performance could be seen as the availability of full-text articles is by far and away the limiting factor. Unfortunately this is not a technical challenge that can be solved by a little creative thinking, rather, it is a challenge in reconciling the different needs and motivations of a number of disparate groups. There has been real progress in recent years with governments realising the worth of text-mining and introducing new legislature legitimising its use, and a willingness by publishers to compromise on access to their intellectual property.

24 Exploiting LiMPET

LiMPET has been integrated with other tools. The output SBML file produced by the tool can be read by the metabolic pathway analysis tool Metingear [152] developed at the EBI. In addition to displaying the basic extracted reactions, Metingear can also display links to the source articles and the extraction and relevance scores. Using Metingear, data obtained through text-mining can easily be integrated with data from other sources.

Throughout this project I have worked with scientists at Unilever with the aim of integrating LiMPET with their systems — specifically the pipeline software Pipeline Pilot from Accelrys. Pipeline Pilot allows researchers with limited programming experience to build pipelines

from a host of modular components. I have written a wrapper for LiMPET allowing the tool to be incorporated into pipelines to allow the mining of internal documents.

While the integration of LiMPET with these third-party tools can allow individual researchers to carry out their own text-mining analysis of metabolic pathways, a more compelling use-case would be its use in assisting database curation. In MetaCyc, data for specific metabolic reactions can originate from a limited number of sources. Using LiMPET, curators would be able to find extra evidence with the only manual effort being to check that the specific source articles for the returned reactions.

In Section 2 I described how databases such as BioCyc and KEGG predict metabolic pathways based on genome sequences. While these predictions provide a useful starting point for research, the reliability of these predictions is unknown. LiMPET could be used to find potential evidence for predicted reactions, with no manual effort, that could be displayed to the user (and clearly marked as having been found using text-mining). The use of text-mining would encourage a more collaborative approach to pathway curation. With users checking the reliability of retrieved evidence themselves, they could also report their findings to inform others. Evidence flagged as unreliable by multiple users would be brought to the attention of manual curators who could remove it. Likewise, predicted pathways with reliable evidence could become potential targets of manual curation.

References

[1] Renata Kabiljo, Andrew B Clegg, and Adrian J Shepherd. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10:233, 2009.

[2] Pubmed help - ncbi bookshelf.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.

[4] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. Cath–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1108, Aug 1997.

[5] Ida Schomburg, Antje Chang, Sandra Placzek, Carola Söhngen, Michael Rother, Maren Lang, Cornelia Munaretto, Susanne Ulas, Michael Stelzer, Andreas Grote, Maurice Scheer, and Dietmar Schomburg. Brenda in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in brenda. *Nucleic Acids Res*, 41(Database issue):D764–D772, Jan 2013.

[6] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1):29–34, Jan 1999.

[7] Ron Caspi, Tomer Altman, Joseph M Dale, Kate Dreher, Carol A Fulcher, Fred Gilham, Pallavi Kaipa, Athikkattuvalasu S Karthikeyan, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A Mueller, Suzanne Paley, Liviu Popescu, Anuradha Pujar, Alexander G Shearer, Peifen Zhang, and Peter D Karp. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 38(Database issue):D473–D479, Jan 2010.

[8] Peter McQuilton and FlyBase Consortium. Opportunities for text mining in the flybase genetic literature curation workflow. *Database (Oxford)*, 2012:bas039, 2012.

[9] Peter D Karp, Christos A Ouzounis, Caroline Moore-Kochlacs, Leon Goldovsky, Pallavi Kaipa, Dag Ahrén, Sophia Tsoka, Nikos Darzentas, Victor Kunin, and Nuria Lopez-Bigas. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*, 33(19):6083–6089, 2005.

[10] Suzanne M Paley and Peter D Karp. Evaluation of computational metabolic-pathway predictions for helicobacter pylori. *Bioinformatics*, 18(5):715–724, May 2002.

[11] Adi Mano, Tamir Tuller, Oded Béjà, and Ron Y Pinter. Comparative classification of species and the study of pathway evolution based on the alignment of metabolic pathways. *BMC Bioinformatics*, 11 Suppl 1:S38, 2010.

[12] Elissavet Nikolaou, Ino Agrafioti, Michael Stumpf, Janet Quinn, Ian Stansfield, and Alistair J P Brown. Phylogenetic diversity of stress signalling pathways in fungi. *BMC Evol Biol*, 9:44, 2009.

[13] M. A. Huynen, T. Dandekar, and P. Bork. Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol*, 7(7):281–291, Jul 1999.

[14] Tara A Gianoulis, Jeroen Raes, Prianka V Patel, Robert Bjornson, Jan O Korbel, Ivica Letunic, Takuji Yamada, Alberto Paccanaro, Lars J Jensen, Michael Snyder, Peer Bork, and Mark B Gerstein. Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A*, 106(5):1374–1379, Feb 2009.

[15] E. Huala, A. W. Dickerman, M. Garcia-Hernandez, D. Weems, L. Reiser, F. LaFond, D. Hanley, D. Kiphart, M. Zhuang, W. Huang, L. A. Mueller, D. Bhattacharyya, D. Bhaya, B. W. Sobral, W. Beavis, D. W. Meinke, C. D. Town, C. Somerville, and S. Y. Rhee. The arabidopsis information resource (tair): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*, 29(1):102–105, Jan 2001.

[16] Erika Check Hayden. Popular plant database set to charge users. *Nature*, Aug 2013.

[17] Brenda - the enzyme database products - biobase biological databases.

[18] Maren Lang, Michael Stelzer, and Dietmar Schomburg. Bkm-react, an integrated biochemical reaction database. *BMC Biochem*, 12:42, 2011.

[19] H. P. Luhn. A business intelligence system. *IBM J. Res. & Dev.*, 2(4):314–319, Oct 1958.

[20] Dietrich Rebholz-Schuhmann, Harald Kirsch, and Francisco Couto. Facts from text–is text mining ready to deliver? *PLoS Biol*, 3(2):e65, Feb 2005.

[21] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del Toro, and et al. The mintact project–intact as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):D358–D363, Nov 2013.

[22] Sandra Orchard, Samuel Kerrien, Sara Abbani, Bruno Aranda, Jignesh Bhate, Shelby Bidwell, Alan Bridge, Leonardo Briganti, Fiona Brinkman, Gianni Cesareni, and et al. Protein interaction data curation: the international molecular exchange (imex) consortium. *Nature Methods*, 9(4):345–350, Mar 2012.

[23] Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. Overview of the protein-protein interaction annotation extraction task of biocreative ii. *Genome Biol*, 9 Suppl 2:S4, 2008.

[24] Yusuke Miyao, Kenji Sagae, Rune Saetre, Takuya Matsuzaki, and Jun'ichi Tsujii. Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics*, 25(3):394–400, Feb 2009.

[25] Lawrence Hunter, Zhiyong Lu, James Firby, William A Baumgartner, Helen L Johnson, Philip V Ogren, and K. Bretonnel Cohen. Opendmap: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, 9:78, 2008.

[26] Dietrich Rebholz-Schuhmann, Miguel Arregui, Sylvain Gaudan, Harald Kirsch, and Antonio Jimeno. Text processing through web services: calling whatizit. *Bioinformatics*, 24(2):296–298, Jan 2008.

[27] Martin Krallinger, Miguel Vazquez, Florian Leitner, David Salgado, Andrew Chatr-Aryamontri, Andrew Winter, Livia Perfetto, Leonardo Briganti, Luana Licata, Marta Iannuccelli, Luisa Castagnoli, Gianni Cesareni, Mike Tyers, Gerold Schneider, Fabio Rinaldi, Robert Leaman, Graciela Gonzalez, Sergio Matos, Sun Kim, W John Wilbur, Luis Rocha, Hagit Shatkay, Ashish V. Tendulkar, Shashank Agarwal, Feifan Liu, Xinglong Wang, Rafal Rak, Keith Noto, Charles Elkan, Zhiyong Lu, Rezarta Islamaj Dogan, Jean-Fred Fontaine, Miguel A. Andrade-Navarro, and Alfonso Valencia. The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC Bioinformatics*, 12 Suppl 8:S3, 2011.

[28] D. Kwon, S. Kim, S.-Y. Shin, A. Chatr-aryamontri, and W. J. Wilbur. Assisting manual literature curation for protein-protein interactions using bioqrator. *Database*, 2014(0):bau067, Jan 2014.

[29] Daniel G Jamieson, Martin Gerner, Farzaneh Sarafraz, Goran Nenadic, and David L Robertson. Towards semi-automated curation: using text mining to recreate the hiv-1, human protein interaction database. *Database (Oxford)*, 2012:bas023, 2012.

[30] Robert Leaman and Graciela Gonzalez. Banner: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput*, pages 652–663, 2008.

[31] Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. Complex event extraction at pubmed scale. *Bioinformatics*, 26(12):i382–i390, Jun 2010.

[32] Makoto Miwa, Rune Saetre, Jin-Dong Kim, and Jun'ichi Tsujii. Event extraction with complex event classification using rich features. *J Bioinform Comput Biol*, 8(1):131–146, Feb 2010.

[33] Lishuang Li, Panpan Zhang, Tianfu Zheng, Hongying Zhang, Zhenchao Jiang, and Degen Huang. Integrating semantic information into multiple kernels for protein-protein interaction extraction from biomedical literatures. *PLoS One*, 9(3):e91898, 2014.

[34] Changqin Quan, Meng Wang, and Fuji Ren. An unsupervised text mining method for relation extraction from biomedical literature. *PLoS One*, 9(7):e102039, 2014.

[35] Jin Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado, June 2009. Association for Computational Linguistics.

[36] Christian Blaschke and Alfonso Valencia. The frame-based module of the suiseki information extraction system. *IEEE Intelligent Systems*, 17:14–20, March 2002.

[37] Ivan Iossifov, Michael Krauthammer, Carol Friedman, Vasileios Hatzivassiloglou, Joel S Bader, Kevin P White, and Andrey Rzhetsky. Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, 20(8):1205–1213, May 2004.

[38] Andrey Rzhetsky, Ivan Iossifov, Tomohiro Koike, Michael Krauthammer, Pauline Kra, Mitzi Morris, Hong Yu, Pablo Ariel Duboué, Wubin Weng, W. John Wilbur, Vasileios Hatzivassiloglou, and Carol Friedman. Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J Biomed Inform*, 37(1):43–53, Feb 2004.

[39] Carlos Santos, Daniela Eggle, and David J States. Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics*, 21(8):1653–1658, Apr 2005.

[40] Anton Yuryev, Zufar Mulyukov, Ekaterina Kotelnikova, Sergei Maslov, Sergei Egorov, Alexander Nikitin, Nikolai Daraselia, and Ilya Mazo. Automatic pathway building in biological association networks. *BMC Bioinformatics*, 7:171, 2006.

[41] Byron Marshall, Hua Su, Daniel McDonald, Shauna Eggers, and Hsinchun Chen. Aggregating automatically extracted regulatory pathway relations. *IEEE Trans Inf Technol Biomed*, 10(1):100–108, Jan 2006.

[42] Carlos Rodríguez-Penagos, Heladia Salgado, Irma Martínez-Flores, and Julio Collado-Vides. Automatic reconstruction of a bacterial regulatory network using natural language processing. *BMC Bioinformatics*, 8:293, 2007.

[43] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1, 2005.

[44] Larry Smith, Lorraine K Tanabe, Rie Johnson nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A Struble, Richard J Povinelli, Andreas Vlachos, William A Baumgartner, Lawrence Hunter, Bob Carpenter, Richard

Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Ma na López, Jacinto Mata, and W. John Wilbur. Overview of biocreative ii gene mention recognition. *Genome Biol*, 9 Suppl 2:S2, 2008.

[45] Zhiyong Lu, Hung-Yu Kao, Chih-Hsuan Wei, Minlie Huang, Jingchen Liu, Cheng-Ju Kuo, Chun-Nan Hsu, Richard Tzong-Han Tsai, Hong-Jie Dai, Naoaki Okazaki, Han-Cheol Cho, Martin Gerner, Illes Solt, Shashank Agarwal, Feifan Liu, Dina Vishnyakova, Patrick Ruch, Martin Romacker, Fabio Rinaldi, Sanmitra Bhattacharya, Padmini Srinivasan, Hongfang Liu, Manabu Torii, Sergio Matos, David Campos, Karin Verspoor, Kevin M Livingston, and W. John Wilbur. The gene normalization task in biocreative iii. *BMC Bioinformatics*, 12 Suppl 8:S2, 2011.

[46] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac Symp Biocomput*, pages 505–516, 2000.

[47] Robert Gaizauskas, Kevin Humphreys, and George Demetriou. Information extraction from biological science journal articles: enzyme interactions and protein structures. In *M.G. Hicks (Ed.), Proceedings of the Workshop Chemical Data Analysis in the Large: the Challenge of the Automation Age*, 2001.

[48] Svetlana Novichkova, Sergei Egorov, and Nikolai Daraselia. Medscan, a natural language processing engine for medline abstracts. *Bioinformatics*, 19(13):1699–1706, Sep 2003.

[49] Nikolai Daraselia, Anton Yuryev, Sergei Egorov, Svetalana Novichkova, Alexander Nikitin, and Ilya Mazo. Extracting human protein interactions from medline using a full-sentence parser. *Bioinformatics*, 20(5):604–611, Mar 2004.

[50] Robert Hoffmann, Martin Krallinger, Eduardo Andres, Javier Tamames, Christian Blaschke, and Alfonso Valencia. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE*, 2005(283):pe21, May 2005.

[51] Chikashi Nobata, Paul D Dobson, Syed A Iqbal, Pedro Mendes, Jun'ichi Tsujii, Douglas B Kell, and Sophia Ananiadou. Mining metabolites: extracting the yeast metabolome from the literature. *Metabolomics*, 7(1):94–101, Mar 2011.

[52] Ian Donaldson, Joel Martin, Berry de Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D Bader, Katerina Michalickova, Tony Pawson, and Christopher W V Hogue. Prebind and textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4:11, Mar 2003.

[53] Nikiforos Karamanis, Ian Lewin, Ruth Seal, Rachel Drysdale, and Edward Briscoe. Integrating natural language processing with flybase curation. *Pac Symp Biocomput*, pages 245–256, 2007.

[54] FlyBase Consortium. The flybase database of the drosophila genome projects and community literature. *Nucleic Acids Res*, 31(1):172–175, Jan 2003.

[55] Rainer Winnenburg, Thomas Wächter, Conrad Plake, Andreas Doms, and Michael Schroeder. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief Bioinform*, 9(6):466–478, Nov 2008.

[56] Thomas C Wiegers, Allan Peter Davis, K. Bretonnel Cohen, Lynette Hirschman, and Carolyn J Mattingly. Text mining and manual curation of chemical-gene-disease networks for the comparative toxicogenomics database (ctd). *BMC Bioinformatics*, 10:326, 2009.

[57] K. Bretonnel Cohen, Helen L Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E Hunter. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11:492, 2010.

[58] Antonio Jimeno Yepes and Karin Verspoor. Literature mining of genetic variants for curation: quantifying the importance of supplementary material. *Database (Oxford)*, 2014(0):bau003, 2014.

[59] Funders - about - europe pubmed central.

[60] The copyright and rights in performances (research, education, libraries and archives) regulations 2014.

[61] Richard Van Noorden. Elsevier opens its papers to text-mining. *Nature*, 506(7486):17, Feb 2014.

[62] Crossref text and data mining.

[63] Liber responds to elsevier's text and data mining policy.

[64] Epccopyright vision 2014.

[65] Open letter responding to epc copyright vision paper 2014: Copyright enabled on the network.

[66] Nikiforos Karamanis, Ruth Seal, Ian Lewin, Peter McQuilton, Andreas Vlachos, Caroline Gasperin, Rachel Drysdale, and Ted Briscoe. Natural language processing in aid of flybase curators. *BMC Bioinformatics*, 9:193, 2008.

[67] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[68] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res*, 41(Web Server issue):W518–W522, Jul 2013.

[69] Marc E. Colosimo, Alexander A. Morgan, Alexander S. Yeh, Jeffrey B. Colombe, and Lynette Hirschman. Data preparation and interannotator agreement: Biocreative task 1b. *BMC Bioinformatics*, 6 Suppl 1:S12, 2005.

[70] JA Swets. Effectiveness of information retrieval methods. Technical report, Cambridge, MA: Bolt, Beranek, and Newman, Inc., 1967.

[71] W.John Wilbur. An information measure of retrieval performance. *Information Systems*, 17(4):283–298, Jul 1992.

[72] Hyrum D. Carroll, Maricel G. Kann, Sergey L. Sheetlin, and John L. Spouge. Threshold average precision (tap-k): a measure of retrieval designed for bioinformatics. *Bioinformatics*, 26(14):1708–1713, Jul 2010.

[73] M. Gribskov and N. L. Robinson. Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Comput Chem*, 20(1):25–33, Mar 1996.

[74] Raoul Frijters, Marianne van Vugt, Ruben Smeets, René van Schaik, Jacob de Vlieg, and Wynand Alkema. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol*, 6(9), 2010.

[75] Michael L Sierk and William R Pearson. Sensitivity and selectivity in protein structure comparison. *Protein Sci*, 13(3):773–785, Mar 2004.

[76] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. 2011.

[77] Oasis.

[78] Oasis members approve open standard for accessing unstructured information.

[79] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–513, 2010.

[80] Yoshinobu Kano, William A Baumgartner, Luke McCrohon, Sophia Ananiadou, K. Bretonnel Cohen, Lawrence Hunter, and Jun'ichi Tsujii. U-compare: share and compare text mining tools with uima. *Bioinformatics*, 25(15):1997–1998, Aug 2009.

[81] Opennlp at apache incubator.

[82] Andrew B. Clegg and Adrian J. Shepherd. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics*, 8:24, 2007.

[83] Ekaterina Buyko, Joachim Wermter, Michael Poprat, and Udo Hahn. Automatically adapting an nlp core engine to the biology domain. In *Proceedings of the ISMB 2006 Joint Linking Literature, Information and Knowledge for Biology and the 9th Bio-Ontologies Meeting*, 2006.

[84] J-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. Genia corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1:i180–i182, 2003.

[85] S. Kulick, A. Bies, M. Liberman, M. Mandel, R. Mcdonald, M. Palmer, and A. Schein. Integrated annotation for biomedical information extraction. In *Biolink: Linking Biological Literature, Ontologies and Databases, Proceedings of HLT-NAACL*, pages 61–68, 2004.

[86] Julie lab opennlp models.

[87] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA*, pages 104–107, 2004.

[88] Burr Settles. Abner: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, Jul 2005.

[89] Javier Tamames. Text detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics*, 6 Suppl 1:S10, 2005.

[90] Jörg Hakenberg, Martin Gerner, Maximilian Haeussler, Illés Solt, Conrad Plake, Michael Schroeder, Graciela Gonzalez, Goran Nenadic, and Casey M Bergman. The gnat library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771, Oct 2011.

[91] Corinna Kolarik, Roman Klinger, Christoph M. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. Chemical names: Terminological resources and corpora annotation. In *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*, Marrakech, Morocco, 2008.

[92] Meenakshi Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. A biological named entity recognizer. *Pac Symp Biocomput*, pages 427–438, 2003.

[93] Peter Corbett and Peter Murray-Rust. Highthroughput identification of chemistry in life science texts. In *In Proceedings of the 2nd International Symposium on Computational Life Science (CompLife '06)*, pages 107–118, 2006.

[94] David M Jessop, Sam E Adams, Egon L Willighagen, Lezan Hawizy, and Peter Murray-Rust. Oscar4: a flexible architecture for chemical text-mining. *J Cheminform*, 3(1):41, 2011.

[95] Tim Rocktäschel, Michael Weidlich, and Ulf Leser. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640, Jun 2012.

[96] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Model.*, 28(1):31–36, Feb 1988.

[97] Alan McNaught. The iupac international chemical identifier: Inchi — a new standard for molecular informatics. *Chemistry international*, pages 12–14, 2006.

[98] Saber A. Akhondi, Jan A. Kors, and Sorel Muresan. Consistency of systematic chemical identifiers within and between small-molecule databases. *J Cheminform*, 4(1):35, 2012.

[99] Daniel M Lowe, Peter T Corbett, Peter Murray-Rust, and Robert C Glen. Chemical name to structure: Opsin, an open source solution. *J Chem Inf Model*, 51(3):739–753, Mar 2011.

[100] Yanfang Sun, Hui Gao, Ying-Wei Yang, Anning Wang, Guolin Wu, Yinong Wang, Yunge Fan, and Jianbiao Ma. Layer-by-layer supramolecular assemblies based on linear and star-shaped poly(glycerol methacrylate)s for doxorubicin delivery. *J Biomed Mater Res A*, 101(8):2164–2173, Aug 2013.

[101] Morihiko Hamada, Edakkattuparambil Sidharth Shibu, Tamitake Itoh, Manikantan Syamala Kiran, Shunsuke Nakanishi, Mitsuru Ishikawa, and Vasudevanpillai Biju. Single-molecule photochemical reactions of auger-ionized quantum dots. *Nano Rev*, 2, 2011.

[102] Evan E. Bolton, Yanli Wang, Paul A. Thiessen, and Stephen H. Bryant. Chapter 12 pubchem: Integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry*, pages 217–241, 2008.

[103] P. de Matos, M. Ennis, M. Darsow, M. Guedj, K. Degtyarenko, and R. Apweiler. Chebi — chemical entities of biological interest. Database Summary Paper 646, EMBL Outstation - The European Bioinformatics Institute, 2006.

[104] Astrid Fleischmann, Michael Darsow, Kirill Degtyarenko, Wolfgang Fleischmann, Sinéad Boyce, Kristian B. Axelsen, Amos Bairoch, Dietmar Schomburg, Keith F. Tipton, and Rolf Apweiler. Intenz, the integrated relational enzyme database. *Nucleic Acids Res*, 32(Database issue):D434–D437, Jan 2004.

[105] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280, Jan 2004.

[106] A. Golovin, T. J. Oldfield, J. G. Tate, S. Velankar, G. J. Barton, H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Hussain, J M C. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, A. Pajon, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, G. J. Swaminathan, M. Tagari, S. Tromm, W. Vranken,

and K. Henrick. E-msd: an integrated data resource for bioinformatics. *Nucleic Acids Res*, 32(Database issue):D211–D216, Jan 2004.

[107] John Overington. Chembl. an interview with john overington, team leader, chemogenomics at the european bioinformatics institute outstation of the european molecular biology laboratory (embl-ebi). interview by wendy a. warr. *J Comput Aided Mol Des*, 23(4):195–198, Apr 2009.

[108] Austin R. Mast and Kevin Thiele. The transfer of *Dryandra* r.br. to *Banksia* l.f. (proteaceae). *Australian Systematic Botany*, 20(1):63, 2007.

[109] Drew Koning, Indra Neil Sarkar, and Thomas Moritz. Taxongrab: Extracting taxonomic names from text. *Biodiversity Informatics*, 2:79–82, 2005.

[110] Guido Sautter, Klemens Böhm, and Donat Agosti. A combining approach to find all taxon names (fat) in legacy biosystematics literature. *Biodiversity Informatics*, 3:46–58, 2006.

[111] Martin Gerner, Goran Nenadic, and Casey M Bergman. Linnaeus: a species name identification system for biomedical literature. *BMC Bioinformatics*, 11:85, 2010.

[112] Nona Naderi, Thomas Kappler, Christopher J O Baker, and RenÃ© Witte. Organismtagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729, Oct 2011.

[113] Jung - java universal network/graph framework.

[114] Graphviz | graphviz - graph visualization software.

[115] Gephi makes graphs handly.

[116] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le NovÃšre, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, J. Wang, and S. B. M. L. Forum. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, Mar 2003.

[117] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl Schaefer, Joanne Luciano, Frank Schacherer, Irma Martinez-Flores, Zhenjun Hu, Veronica Jimenez-Jacinto, Geeta Joshi-Tope, Kumaran Kandasamy, Alejandra C Lopez-Fuentes, Huaiyu Mi, Elgar Pichler, Igor Rodchenkov, Andrea Splendiani, Sasha Tkachev, Jeremy Zucker, Gopal Gopinath, Harsha

Rajasimha, Ranjani Ramakrishnan, Imran Shah, Mustafa Syed, Nadia Anwar, OzgÃŒn Babur, Michael Blinov, Erik Brauner, Dan Corwin, Sylva Donaldson, Frank Gibbons, Robert Goldberg, Peter Hornbeck, Augustin Luna, Peter Murray-Rust, Eric Neumann, Oliver Reubenacker, Matthias Samwald, Martijn van Iersel, Sarala Wimalaratne, Keith Allen, Burk Braun, Michelle Whirl-Carrillo, Kei-Hoi Cheung, Kam Dahlquist, Andrew Finney, Marc Gillespie, Elizabeth Glass, Li Gong, Robin Haw, Michael Honig, Olivier Hubaut, David Kane, Shiva Krupa, Martina Kutmon, Julie Leonard, Debbie Marks, David Merberg, Victoria Petri, Alex Pico, Dean Ravenscroft, Liya Ren, Nigam Shah, Margot Sunshine, Rebecca Tang, Ryan Whaley, Stan Letovksy, Kenneth H Buetow, Andrey Rzhetsky, Vincent Schachter, Bruno S Sobral, Ugur Dogrusoz, Shannon McWeeney, Mirit Aladjem, Ewan Birney, Julio Collado-Vides, Susumu Goto, Michael Hucka, Nicolas Le NovÃšre, Natalia Maltsev, Akhilesh Pandey, Paul Thomas, Edgar Wingender, Peter D Karp, Chris Sander, and Gary D Bader. The biopax community standard for pathway data sharing. *Nat Biotechnol*, 28(9):935–942, Sep 2010.

[118] Jan Czarnecki, Irene Nobeli, Adrian M. Smith, and Adrian J. Shepherd. A text-mining system for extracting metabolic reactions from full-text articles. *BMC Bioinformatics*, 13:172, 2012.

[119] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, Feb 2001.

[120] Colin R. Batchelor and Peter T. Corbett. Semantic enrichment of journal articles using chemical named entity recognition. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 45–48, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[121] J. W. Patrick and N. Lee. Purification and properties of an l-arabinose isomerase from escherichia coli. *J Biol Chem*, 243(16):4312–4318, Aug 1968.

[122] Porter stemming algorithm implementations.

[123] MF Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[124] Edwin Webb. *Enzyme Nomenclature 1992*. Published for the International Union of Biochemistry and Molecular Biology by Academic Press, San Diego, 1992.

[125] Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10, 2008.

[126] S. Iuchi, M. Kobayashi, T. Taji, M. Naramoto, M. Seki, T. Kato, S. Tabata, Y. Kakubari, K. Yamaguchi-Shinozaki, and K. Shinozaki. Regulation of drought tolerance by gene manipulation of 9-cis-epoxycarotenoid dioxygenase, a key enzyme in abscisic acid biosynthesis in arabidopsis. *Plant J*, 27(4):325–333, Aug 2001.

[127] Parantu K Shah, Carolina Perez-Iratxeta, Peer Bork, and Miguel A Andrade. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 4:20, May 2003.

[128] R. Kabiljo and Adrian J. Shepherd. Protein name tagging in the immunological domain. In *Proceedings of the Third Symposium on Semantic Mining in Biomedicine*, Turku, Finland, September 2008.

[129] Alexander Yeh, Alexander Morgan, Marc Colosimo, and Lynette Hirschman. Biocreative task 1a: gene mention finding evaluation. *BMC Bioinformatics*, 6 Suppl 1:S2, 2005.

[130] R. Kabiljo, D. Stoycheva, and A. J. Shepherd. Prospectome: a new tagged corpus for protein named entity recognition. In *Proceedings of The ISMB BioLINK, Special Interest Group on Text Data Mining*, pages 24–27, 2007.

[131] Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artif Intell Med*, 33(2):139–155, Feb 2005.

[132] C. Nédellec. Learning language in logic - genic interaction extraction challenge. In *Proceedings of the Learning Language in Logic 2005 Workshop at the International Conference on Machine Learning*, 2005.

[133] Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relex–relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, Feb 2007.

[134] Rune Sætre, Kenji Sagae, and Jun ichi Tsujii. Syntactic features for protein-protein interaction extraction. In *LBM (Short Papers)'07*, 2007.

[135] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8:50, 2007.

[136] J. D. Wren and H. R. Garner. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods Inf Med*, 41(5):426–434, 2002.

[137] Claudia Catalanotti, Wenqiang Yang, Matthew C. Posewitz, and Arthur R. Grossman. Fermentation metabolism and its evolution in algae. *Front Plant Sci*, 4:150, 2013.

[138] Maurice Scheer, Andreas Grote, Antje Chang, Ida Schomburg, Cornelia Munaretto, Michael Rother, Carola SÃ¶hngen, Michael Stelzer, Juliane Thiele, and Dietmar Schomburg. Brenda, the enzyme information system in 2011. *Nucleic Acids Res*, 39(Database issue):D670–D676, Jan 2011.

[139] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297, Jul 1945.

[140] E. W. Nester and A. L. Montoya. An enzyme common to histidine and aromatic amino acid biosynthesis in bacillus subtilis. *J Bacteriol*, 126(2):699–705, May 1976.

[141] Cheng-Ju Kuo, Maurice H T. Ling, and Chun-Nan Hsu. Soft tagging of overlapping high confidence gene mention variants for cross-species full-text gene normalization. *BMC Bioinformatics*, 12 Suppl 8:S6, 2011.

[142] Minlie Huang, Jingchen Liu, and Xiaoyan Zhu. Genetukit: a software for document-level gene normalization. *Bioinformatics*, 27(7):1032–1033, Apr 2011.

[143] Karen Yook, Todd W. Harris, Tamberlyn Bieri, Abigail Cabunoc, Juancarlos Chan, Wen J. Chen, Paul Davis, Norie de la Cruz, Adrian Duong, Ruihua Fang, Uma Ganesan, Christian Grove, Kevin Howe, Snehalata Kadam, Ranjana Kishore, Raymond Lee, Yuling Li, Hans-Michael Muller, Cecilia Nakamura, Bill Nash, Philip Ozersky, Michael Paulini, Daniela Raciti, Arun Rangarajan, Gary Schindelman, Xiaoqi Shi, Erich M. Schwarz, Mary Ann Tuli, Kimberly Van Auken, Daniel Wang, Xiaodong Wang, Gary Williams, Jonathan Hodgkin, Matthew Berriman, Richard Durbin, Paul Kersey, John Spieth, Lincoln Stein, and Paul W. Sternberg. Wormbase 2012: more genomes, more data, new website. *Nucleic Acids Res*, 40(Database issue):D735–D741, Jan 2012.

[144] Kimberly Van Auken, Petra Fey, Tanya Z. Berardini, Robert Dodson, Laurel Cooper, Donghui Li, Juancarlos Chan, Yuling Li, Siddhartha Basu, Hans-Michael Muller, Rex Chisholm, Eva Huala, Paul W. Sternberg, and WormBase Consortium . Text mining in the biocuration workflow: applications for literature curation at wormbase, dictybase and tair. *Database (Oxford)*, 2012:bas040, 2012.

[145] Hans-Michael Müller, Eimear E. Kenny, and Paul W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, Nov 2004.

[146] Baoping Ling, Min Sun, Siwei Bi, Zhihong Jing, and Yongjun Liu. Molecular dynamics simulations of the coenzyme induced conformational changes of mycobacterium tuberculosis l-alanine dehydrogenase. *J Mol Graph Model*, 35:1–10, May 2012.

[147] Shalini Saxena, Parthiban Brindha Devi, Vijay Soni, Perumal Yogeeswari, and Dharmarajan Sriram. Identification of novel inhibitors against mycobacterium tuberculosis l-alanine dehydrogenase (mtb-aladh) through structure-based virtual screening. *J Mol Graph Model*, 47:37–43, Feb 2014.

[148] Michelle M. Giffin, Lucia Modesti, Ronald W. Raab, Lawrence G. Wayne, and Charles D. Sohaskey. ald of mycobacterium tuberculosis encodes both the alanine dehydrogenase and the putative glycine dehydrogenase. *J Bacteriol*, 194(5):1045–1054, Mar 2012.

[149] Veeraraghavan Usha, Adrian J. Lloyd, Andrew L. Lovering, and Gurdyal S. Besra. Structure and function of mycobacterium tuberculosis meso-diaminopimelic acid (dap) biosynthetic enzymes. *FEMS Microbiol Lett*, 330(1):10–16, May 2012.

[150] Yusuke Shinyama and Satoshi Sekine. Preemptive information extraction using unrestricted relation discovery. In *In Proc. HLT-NAACL-2006*, 2006.

[151] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data, 2009.

[152] John W May, A. Gordon James, and Christoph Steinbeck. Metingear: a development environment for annotating genome-scale metabolic models. *Bioinformatics*, 29(17):2213–2215, Sep 2013.

Part VII

Appendices

25 Appendix I

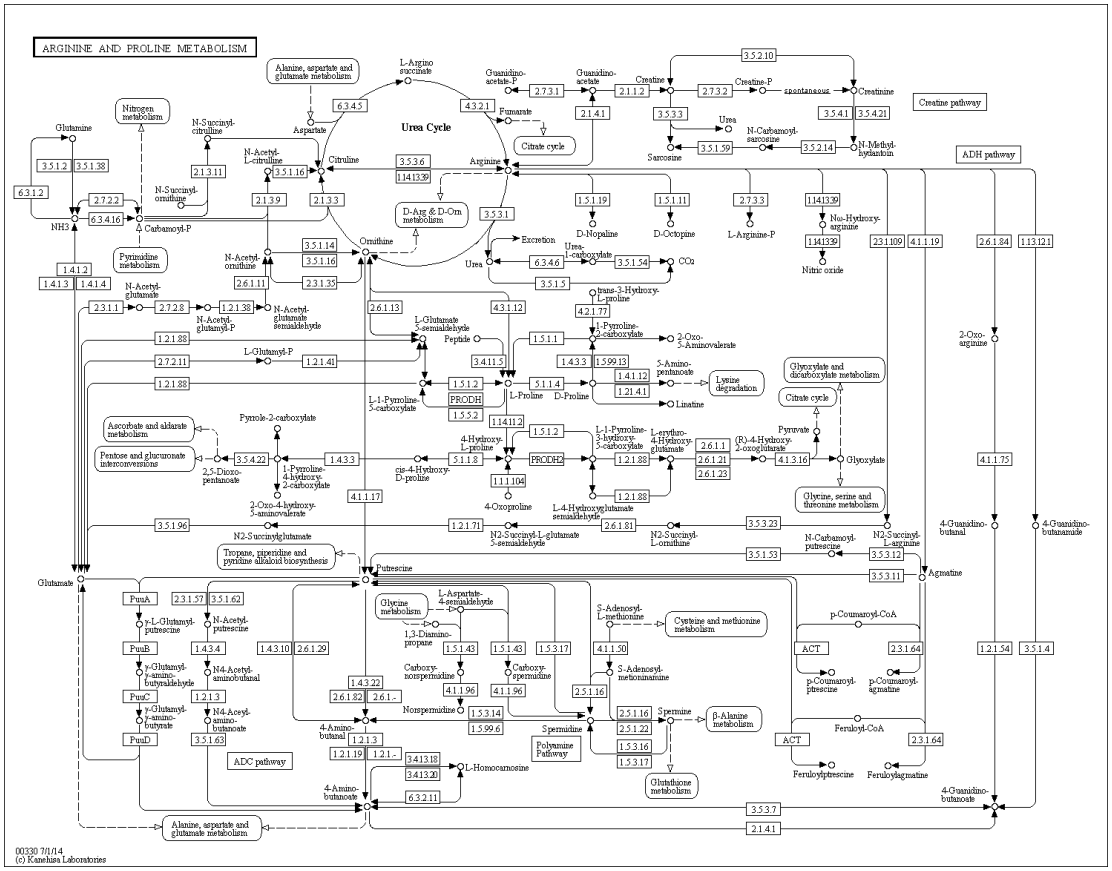

Figure 16: The KEGG network “arginine and proline metabolism”.

26 Appendix II

To allow metabolites to be looked up rapidly in the offline ChEBI database, name variants were pregenerated and indexed. Consider the small molecule *aldehydo-D-glucose 6-phosphate(2−)*. The following variants were generated:

1. Remove round and square brackets with their content.
   aldehydo-D-glucose 6-phosphate
2. Remove stereochemistry identifiers.
   aldehydo-glucose 6-phosphate
3. Remove all whitespace.
   aldehydo-glucose6-phosphate
4. Remove non-word characters (the set of word characters contains the 26 letters, 10 numbers and underscore).
   aldehydoglucose6phosphate
5. Remove any non-letters.
   aldehydoglucosephosphate

Small molecule entities recognised by OSCAR4 undergo the same variant generation. Each variant is used to query the database in turn until a match is found. Consider the extracted entity *aldehydo glucose 6-phosphate*. The following variants corresponding to the pregenerated variants are generated:

1. *aldehydo glucose 6-phosphate* - no match
2. *aldehydo glucose 6-phosphate* - no match
3. *aldehydoglucose6-phosphate* - no match
4. *aldehydoglucose6-phosphate* - matched

27 Appendix III

The manually curated table below was used to identify *currency* molecules. Care was taken to only include small molecules that would be considered *currency* molecules in the vast majority of cases.

Name	InChI
NAD+	InChI=1S/C21H27N7O14P2/c22-17-12-19(25-7-24-17)28(8-26-12)21-16(32)14(30)11(41-21)6-39-44(36,37)42-43(34,35)38-5-10-13(29)15(31)20(40-10)27-3-1-2-9(4-27)18(23)33/h1-4,7-8,10-11,13-16,20-21,29-32H,5-6H2,(H5-,22,23,24,25,33,34,35,36,37)
NADH	InChI=1S/C21H29N7O14P2/c22-17-12-19(25-7-24-17)28(8-26-12)21-16(32)14(30)11(41-21)6-39-44(36,37)42-43(34,35)38-5-10-13(29)15(31)20(40-10)27-3-1-2-9(4-27)18(23)33/h1,3-4,7-8,10-11,13-16,20-21,29-32H,2,5-6H2,(H2,23,33)(H,34,35)(H,36,37)(H2,22,24,25)
NADP+	InChI=1S/C21H28N7O17P3/c22-17-12-19(25-7-24-17)28(8-26-12)21-16(44-46(33,34)35)14(30)11(43-21)6-41-48(38,39)45-47(36,37)40-5-10-13(29)15(31)20(42-10)27-3-1-2-9(4-27)18(23)32/h1-4,7-8,10-11,13-16,20-21,29-31H,5-6H2,(H7-,22,23,24,25,32,33,34,35,36,37,38,39)
NADPH	InChI=1S/C21H30N7O17P3/c22-17-12-19(25-7-24-17)28(8-26-12)21-16(44-46(33,34)35)14(30)11(43-21)6-41-48(38,39)45-47(36,37)40-5-10-13(29)15(31)20(42-10)27-3-1-2-9(4-27)18(23)32/h1,3-4,7-8,10-11,13-16,20-21,29-31H,2,5-6H2,(H2,23,32)(H,36,37)(H,38,39)(H2,22,24,25)(H2,33,34,35)
ATP	InChI=1S/C10H16N5O13P3/c11-8-5-9(13-2-12-8)15(3-14-5)10-7(17)6(16)4(26-10)1-25-30(21,22)28-31(23,24)27-29(18,19)20/h2-4,6-7,10,16-17H,1H2,(H,21,22)(H,23,24)(H2,11,12,13)(H2,18,19,20)
ADP	InChI=1S/C10H15N5O10P2/c11-8-5-9(13-2-12-8)15(3-14-5)10-7(17)6(16)4(24-10)1-23-27(21,22)25-26(18,19)20/h2-4,6-7,10,16-17H,1H2,(H,21,22)(H2,11,12,13)(H2,18,19,20)
AMP	InChI=1S/C10H14N5O7P/c11-8-5-9(13-2-12-8)15(3-14-5)10-7(17)6(16)4(22-10)1-21-23(18,19)20/h2-4,6-7,10,16-17H,1H2,(H2,11,12,13)(H2,18,19,20)
C	InChI=1S/C
O	InChI=1S/O
N	InChI=1S/N
H	InChI=1S/H
CO_2	InChI=1S/CO2/c2-1-3
H_2O	InChI=1S/H2O/h1H2

28 Appendix IV

28.1 Reaction word stems

- to
- convers
- convert
- transfer
- add
- incorpor
- transform
- isomeris
- isomeriz
- isomer
- coupl
- cycliz
- cyclis
- cyclis
- cleav
- dimer
- dimeris
- trimer
- trimeris
- condens
- aromat
- hydrat
- dearomat
- decarboxyl
- dehydrogen
- reduc
- reduct
- oxidis

- oxidiz
- oxid
- dismut
- transhydrogen
- peroxid
- peroxidis
- peroxidiz
- de-epoxid
- hydrogen
- dioxygen
- lipoxygen
- monooxygen
- oxygen
- epoxid
- hydroxyl
- transhydroxyl
- demethyl
- desatur
- ferroxid
- dehalogen
- deiodin
- methyl
- N-methyl
- hydroxymethyl
- formyl
- aminomethyl
- formimin
- carboxyl
- carbamoyl
- amidin
- transketol
- transadol

- acyl
- succinyl
- palmitoyl
- coumaroyl
- acetyl
- arachidonoyl
- benzoyl
- galloyl
- sinapoyl
- tigloyl
- tetradecanoyl
- hydroxycinnamoyl
- feruloyl
- malonyl
- mycolyl
- dihydroxycinnamoyl
- piperoyl
- trimethyltridecanoyl
- myristoyl
- methylpropanoyl
- thiol
- glutamyl
- lysyl
- cyclis
- oxidocyclis
- oxidocycl
- leucyl
- aspartyl
- arginyl
- glutamyl
- alanyl
- glutamylcysteinyl

- phosphoryl
- dextran
- sucrat
- glucano
- galactosyl
- glucosyl
- mannosyl
- acetylglucosaminyl
- galactosyl
- abequosyl
- fucosyl
- acetylglucosaminyl
- acetylgalactosaminyl
- glucuronosyl
- fructosyl
- glycosyl
- glyco
- rhamnosyl
- acetylmannosaminouronosyl
- glucuronosyl
- glucosaminyl
- ribosyl
- deoxyribsyl
- phosphoribosyl
- diphosphoryl
- phosphoribosyl
- apiosyl
- xylosyl
- dihydrostreptosyl
- arabinosyl
- sialyl
- dimethylallyl

- pyridinyl
- sulfuryl
- adenosyl
- carboxyvinyl
- isopentenyl
- geranyl
- octaprenyl
- polyprenyl
- aminocarboxypropyl
- hexaprenyl
- farnesyl
- decaprenyl
- pentaprenyl
- nonaprenyl
- geranylgeranyl
- aminocarboxypropyl
- aminocarboxyetyl
- sulfhydryl
- transamin
- oximin
- purin
- adenyl
- adenylyl
- nucleotidyl
- uridylyl
- guanylyl
- cytidylyl
- thymidylyl
- phosphoryl
- diphosphoryl
- sulfur
- hydrol

- hydr
- hydrolis
- hydrolys
- hydroliz
- hydrolyz
- hydrolysi
- hydrolyzi
- dehydr
- dephosphoryl
- alkyl

28.2 Production word stems

- from
- produc
- product
- yield
- generat
- creat
- construct
- form
- format
- make
- manufactur
- return
- give
- synthes
- synthesi
- synthesis
- biosynthes
- biosynthesi
- biosynthesis
- result

28.3 Scoring locations

The following list describes the locations of entities and reaction and production keywords that score +0.3 for each entity placement and +2 for each keyword placement, where E = enzyme, S = adjacent substrates, P = adjacent products, Rw = reaction keyword and Pw = production keyword. A keyword in any other location will score -1.

- E - Rw - S - Rw - Pw - P
- E - Pw - P - Pw - Rw -S
- Rw - S - Rw - E - Pw - P
- Pw - P - Pw - E - Rw - S
- Rw - S - Rw - Pw - P - E
- Rw - S - Rw - Pw - P
- Pw - P - Pw - Rw - S

In a list of substrates or products the keyword and between the last two metabolites will score +2, but will score -1 if found between any two other entities in the list:

S1 - S2 ... Sn-1 - AND - Sn

29 Appendix V

Reaction	Extraction	Relevance
(E,E,E,)-geranylgeranyl diphosphate -> ent-kaurene	0.982	1
copalyl pyrophosphate + (E,E,E,)-geranylgeranyl diphosphate -> ent-kaurene	0.932	1
(E,E,E,)-geranylgeranyl diphosphate -> copalyl diphosphate	0.995	0.888799
copalyl diphosphate -> ent-kaurene	0.932	0.865598
pyrophosphate -> ent-kaurene	0.752	0.583407
Chlorophyll -> phytol	0.969	0.196073
anylgeranyl pyrophosphate -> copalyl pyrophosphate	0.9132	0.15437
phytol -> phytyl diphosphate + phytyl-phosphate	0.9132	0.137249
ent-kaurene -> ent-kaurenoic acid	0.9132	0.134339
isopentenyl diphosphate + dimethylallyl diphosphate -> geranyl diphosphate	0.932	0.086386
(E,E,E,)-geranylgeranyl diphosphate -> phytoene	0.752	0.076126
copalyl pyrophosphate -> ent-kaurene	0.8884	0.058275
GA12 -> GA9	0.932	0.05665
(E,E,E,)-geranylgeranyl diphosphate -> phytyl diphosphate	0.752	0.041204
(E,E,E,)-geranylgeranyl diphosphate -> tocotrienol	0.752	0.041204
dioxygenase -> tocotrienols	0.752	0.041204
phytol -> phytyl diphosphate	0.752	0.041204
(E,E,E,)-geranylgeranyl diphosphate -> GA	0.752	0.040266
phytol -> Chlorophyll	0.8884	0.039209
gibberellin -> ent-kaurene	0.752	0.039057
farnesyl diphosphate -> squalene	0.8884	0.033427
GA1 -> amino acids	0.752	0.032843
isopentenyl diphosphate -> pyrophosphate	0.932	0.028033
isopentenyl diphosphate + pyrophosphate -> polyprenyl diphosphates	0.752	0.028033
sulfated -> adenosine 5´-phosphosulfate	0.8884	0.022457
...		

Table 14: The truncated list of reactions extracted when using the pathway "ent-kaurene biosynthesis II" as a seed to extract reactions in *Arabidopsis thaliana*. An extraction score threshold of 0.75 was applied and reactions were ordered by relevance score. The green cells show the reactions corresponding to the pathway "ent-kaurene biosynthesis I". The extraction and relevance scores for a specific reaction correspond to the highest scores achieved by any metabolite in the reaction.

30 Appendix VI

Part IV describes the evaluation of LiMPET using sets of alternative pathways from MetaCyc. Using one pathway as a seed, LiMPET was tasked with extracting the corresponding alternative pathway in an organism known to host it. Two examples comparing the seed pathway, target pathway and the pathway extracted by LiMPET follow.

Extracted pathways are drawn as bipartite networks, with separate reaction nodes (blue squares) and metabolite nodes (pink circles). The thickness of connecting arrows correlate with extraction scores while the colour reflects relevance scores (with blue equalling a relevance of 0 and red equalling a relevance of 1). Extracted pathways are laid out automatically by Cytoscape (using the yFiles algorithm *Organic*).

30.1 *Ent*-kaurene biosynthesis

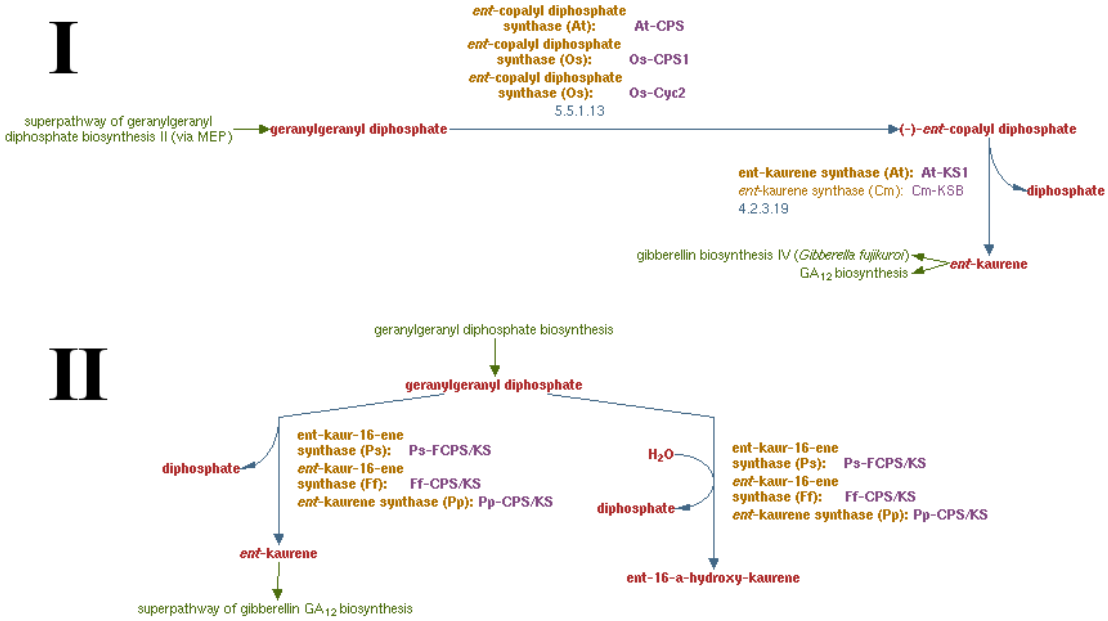

Figure 17: The pathways "*ent*-kaurene biosynthesis" I and II from MetaCyc showing two routes between *geranylgeranyl diphosphate* and *ent-kaurene*. Pathway II was used as a seed pathway to discover the corresponding pathway (I) in *Arabidopsis thaliana* (see Figure 18).

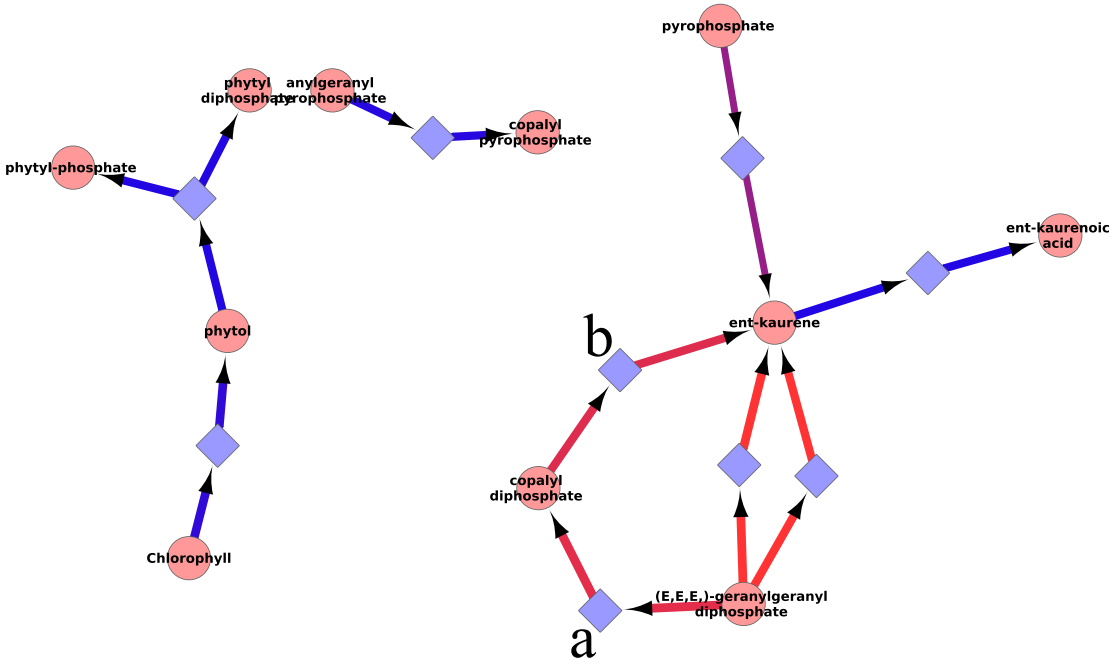

Figure 18: The extracted network when using the pathway "*ent*-kaurene biosynthesis I" as a seed to discover the corresponding pathway in *Arabidopsis thaliana*. Extraction and relevance thresholds were applied. Reactions a and b correspond to the two reactions of the pathway "ent-kaurene biosynthesis II" (see Figure 17).
The reactions directly linking *(E,E,E,)-geranylgeranyl diphosphate* to *ent-kaurene* are assigned a high relevance (shown by their red colour) as they are found in the seed pathway. This reaction appears to be present twice, but one reaction contained a side metabolite which achieved a low relevance score and so is not visible.

30.2 Pyruvate fermentation to ethanol

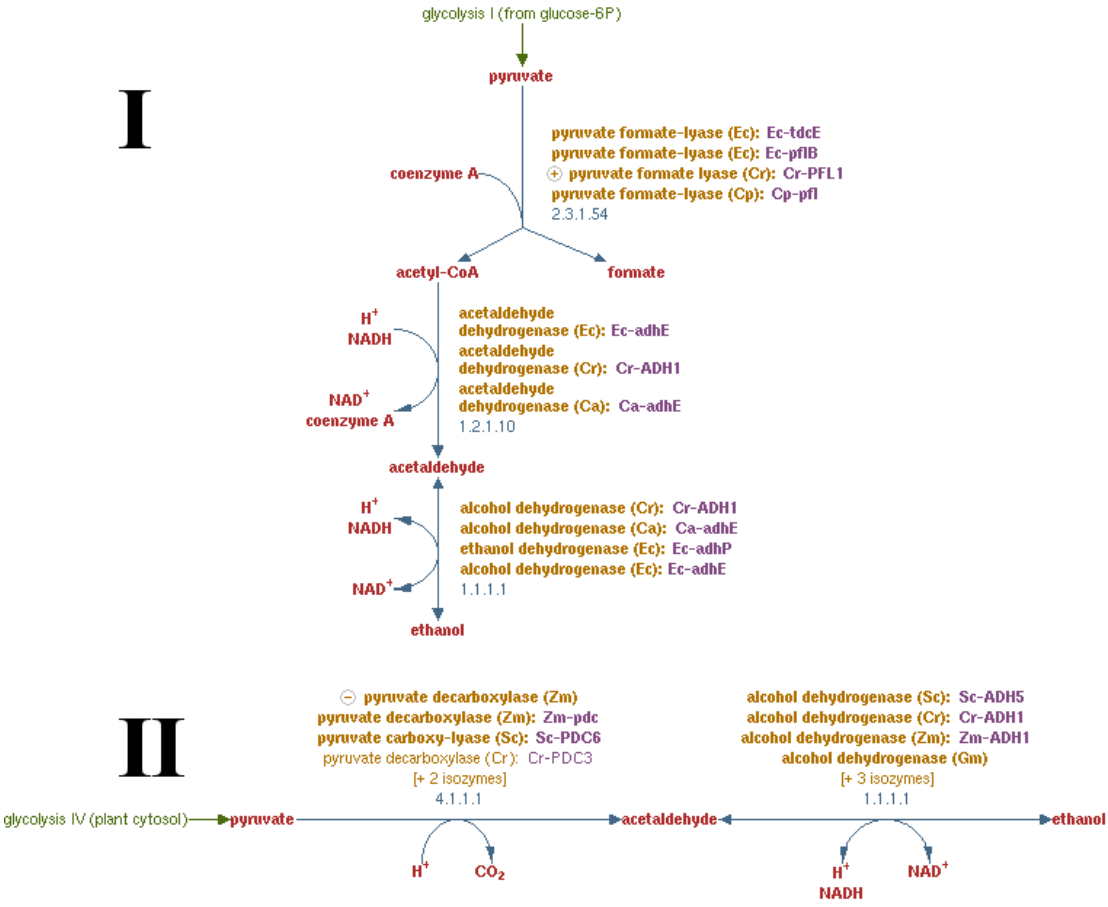

Figure 19: The pathways *"pyruvate fermentation to ethanol"* I and II from MetaCyc showing two routes between *pyruvate* and *ethanol*. Pathway I was used as a seed pathway to discover the corresponding pathway (II) in *Zea mays* (see Figure 20).

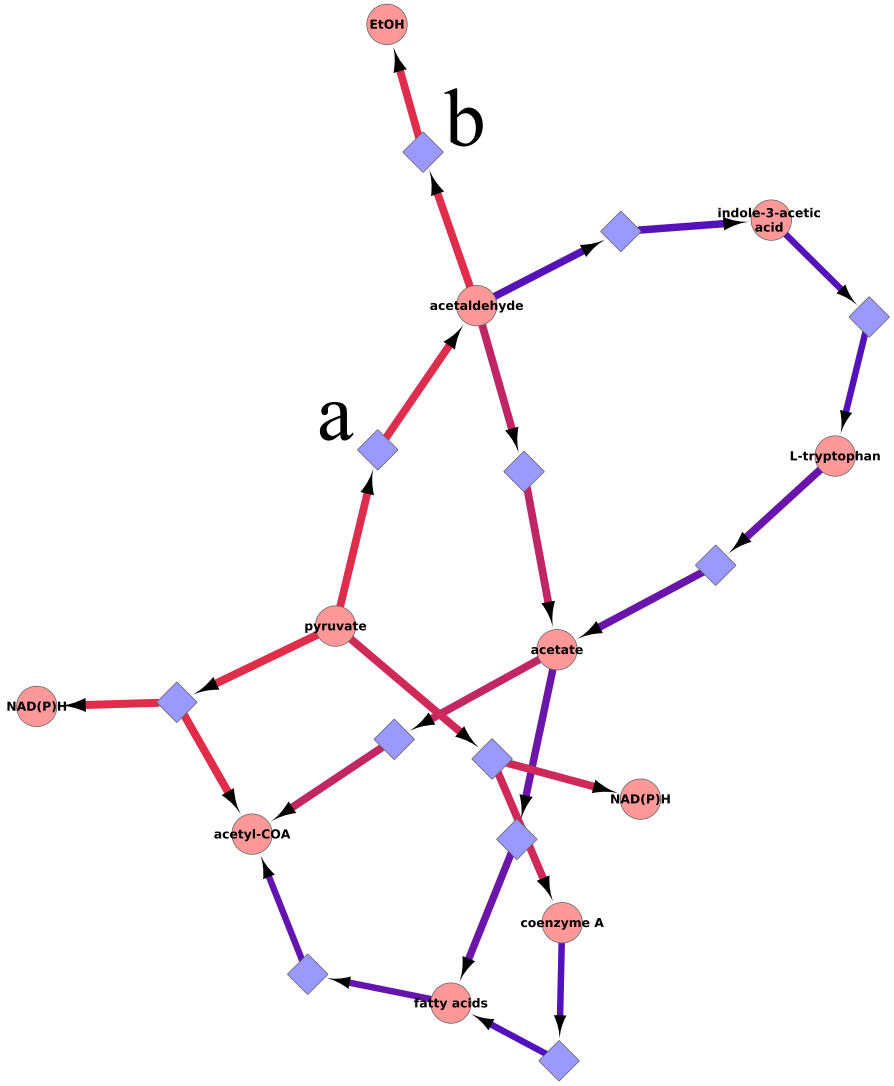

Figure 20: The extracted network when using the pathway "pyruvate fermentation to ethanol I" as a seed to discover the corresponding pathway in *Zea mays*. Extraction and relevance thresholds were applied. Reactions a and b correspond to the two reactions of the pathway "pyruvate fermentation to ethanol II" (see Figure 19).